UTEC-CSc-78-073

Semi-Annual Technical Report

NOISE SUPPRESSION METHODS FOR ROBUST SPEECH PROCESSING

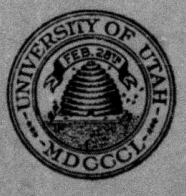
Contractor:	University of Utah
Effective Date:	1 October 1976
Expiration Date:	30 September 1978
Reporting Period:	1 October 1977 – 31 March 1978

Principal Investigator: Dr. Steven F. Boll

Telephone: (801) 581-8224

Sponsored by
Defense Advanced Research Projects Agency (DoD)
ARPA Order No. 3301
Monitored by Naval Research Laboratory
Under Contract No. N00173-77-C-0041

April 1978



SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: UTEC-CSc-78-073	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle): Noise Suppression Methods for Robust Speech Processing.	5. TYPE OF REPORT & PERIOD COVERED: Semi-Annual Technical Rept. 1 Oct 1977-31 Mar 1978
6. PERFORMING ORG. REPORT NUMBER	
7. AUTHOR(s): Steven F. Boll, Dennis Pulsipher, William Done, Ben Cox, Jim Kajiya	8. CONTRACT OR GRANT NUMBER(s): N00143-77-C-0041, ARPA Order-3301
9. PERFORMING ORGANIZATION NAME AND ADDRESS: University of Utah, Computer Science Department, Salt Lake City, Utah 84112	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: Project: 76-RPA-3301
11. CONTROLLING OFFICE NAME AND ADDRESS: Defense Advanced Research Project Agency (DoD), 1400 Wilson Boulevard, Washington, D.C. 22209	12. REPORT DATE: April 1978
	13. NUMBER OF PAGES: 106
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): Naval Research Laboratory, 4555 Overlook Avenue, S.W., Mail Code 2415-A.M. 112p.	15. SECURITY CLASS. (of this report): Unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

This document has been approved for public release and sale; its distribution is unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Digital noise suppression; Linear Predictive Coding; Narrow band coded speech; Adaptive noise cancellation; Weiner filtering; Power spectrum; Autocorrelation, Spectral Averaging for Bias Estimation and Removal (SABER); Widrow-Hoff LMS Algorithm; Pole-zero modeling, Estimation; Constant Q Transform; Non-Parametric, Speech Activity Detector.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Robust speech processing in practical operating environments requires effective environmental and processor noise suppression. This report describes the technical findings and accomplishments during this reporting period for the research program funded to develop real time, compressed speech analysis-synthesis algorithms whose performance is invariant under signal contamination. Fulfillment of this requirement is necessary to insure reliable secure compressed speech transmission within realistic

(over)

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

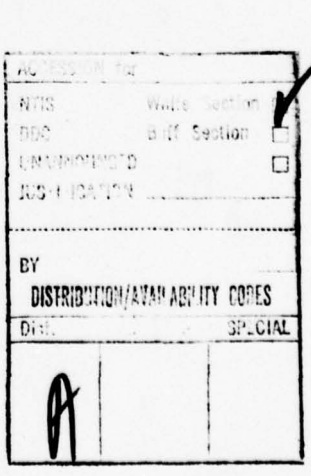
404 949

20. ABSTRACT con't.

military command and control environments. Overall contributions resulting from this research program include the understanding of how environmental noise degrades narrow band, coded speech; development of appropriate real time noise suppression algorithms; and development of speech parameter identification methods that consider signal contamination as a fundamental element in the estimation process. This report describes the current research and results in the areas of noise suppression using the SABER algorithm, dual input adaptive noise cancellation using the LMS algorithm, pole-zero parameter estimation, nonparametric-rank order statistics applications to Robust Speech activity detection, and spectral analysis and synthesis using the constant-Q transform.

TABLE OF CONTENTS

		Page
I.	D D Form 1473	
II.	List of Figures	ii
III.	REPORT SUMMARY	
	Section I Summary of Program for Reporting Period	1
IV.	RESEARCH ACTIVITIES	
	Suppression of Noise in Speech Using the SABER Method	8
	Steven F. Boll	
	A Summary of Recent Experiments Applying Adaptive Noise Cancellation Techniques to Audio Signals	31
	Dennis Pulsipher	
	Estimation of the Parameters of an Autoregressive Moving-Average Process in the Presence of Noise	38
	William J. Done	
	Nonparametric-Rank Order Statistics Applications to Robust Speech Activity Detection	55
	Benjamin V. Cox	
	The Constant-Q Transform	93
	Jim Kajiya	



LIST OF FIGURES

Figure	Page
Data Segmentation	17
Input-Output Relations	21
Block Diagram	22
Table--DRT Scores for Single Speaker	23
Table--Quality Scores from DRT Data	24
Block Diagram of System Identification Model Used in Mode 1 Iterative Procedure of Estimating Parameters	40
10 Pole, 2 Zero Model Spectrum to be Identified	44
Output of Model with an Input	44
Model Spectrum	45
Output of Model with an Inpulse Train Input	45
Real Part of the Complex Cepstrum of x(k)	47
Output of Model with an Inpulse Input	47
Spectrum of x(k) Produced by an Inpulse Train Input to the Model	48
Model Spectrum	48
Output of Model with Noise	49
Model Spectrum	50
Model Spectrum	50
Block Diagram of Signal Classification Method	62
Partitioning of the Speech Spectrum into Four Contiguous Bands that Contribute Equally to Articulation Index. The Frequency Range is 200 to 3200 Hz	64

LIST OF FIGURES Continued

Figure	Page
Real Speech and Theoretical Gamma and Laplace Probability Densities.	68
Table--Recognition Rate for the Simultaneous Decision Procedure for all Seven Words	82
Table--Recognition Rate for the K-W Decision Procedure	83

Section I

Summary of Program for

Reporting Period

Program Objectives

To develop practical, low cost, real time methods for suppressing noise which has been acoustically added to speech.

To demonstrate that through the incorporation of the noise suppression methods, speech can be effectively analysed for narrow band digital transmission in practical operating environments.

Summary of Tasks and Results

Introduction

This semi-annual technical report describes the current status in five research areas for the period 1 October 1977 through 31 March 1978.

SUPPRESSION OF NOISE IN SPEECH USING THE SABER METHOD

Steven F. Boll

A stand alone noise suppression algorithm is described for reducing the spectral effects of acoustically added noise in speech. A fundamental result is developed which shows that the spectral magnitude of speech plus noise can be effectively approximated as the sum of magnitudes of speech and noise. Using this simple phase independent additive model, the noise bias present in the short time spectrum is reduced by subtracting off the expected noise spectrum calculated during nonspeech activity. After bias removal, the time waveform is recalculated from the modified magnitude and saved phase. This Spectral Averaging for Bias Estimation and Removal, or SABER method requires only one FFT per time window for analysis and synthesis.

A SUMMARY OF RECENT EXPERIMENTS APPLYING ADAPTIVE NOISE CANCELLATION TECHNIQUES TO AUDIO SIGNALS

Dennis Pulsipher

A dual input noise cancellation technique for audio signals was presented in a semi-annual report a year ago. The philosophy behind the technique was quite different from that of traditional techniques. Instead of estimating the desired signal directly, the technique attempted to estimate the noise directly and obtained a signal estimate by subtracting the noise estimate from the noisy signal.

The experiments which had been performed at that time used synthetic data and demonstrated great potential for the technique. In the last semi-annual report initial experiments in a real environment were described. A description of experiments that have followed and some of the questions they have raised comprises the body of this report.

During these experiments it became obvious that many facets of the noise cancellation problem are yet to be understood. Techniques dealing with filter inversion are being investigated to better understand the problems

involved. Investigations are also underway to improve convergence of channel estimates when frequency bands of low energy are contained in the reference noise samples. Even if these investigations are fruitless, however, noise cancellation now appears to be a worthwhile approach to signal restoration in acoustically hostile environments.

Estimation of the Parameters of an Autoregressive Moving-Average Process In the Presence of Noise

William J. Done

The previous report on this project presented the details for the autoregressive moving-average (ARMA) process generated by adding white noise to an autoregressive (AR) process. That report stressed the problems that are inherent in estimating the parameters of the resulting ARMA process. Part of this estimation problem lies in the validity of this model for a given application. A major part of the difficulty, however, lies in developing estimation procedures for ARMA processes, regardless of the source of that process. The primary effort in this project since the last report has been the investigation of various methods that might be used to estimate the autoregressive and moving-average coefficients of an ARMA process from data generated by that process. Three methods have been implemented for evaluation.

Nonparametric-Rank Order Statistics Applications

To Robust Speech Activity Detection

Benjamin V. Cox

This report describes a theoretical and experimental investigation for detecting the presence of speech in wideband noise. An algorithm for making the silence-speech decision is described. This algorithm is based on a nonparametric statistical signal-detection scheme that does not require a training set of data and maintains a constant false alarm rate for a broad class of noise inputs. The nonparametric decision procedure is the multiple use of the two-sample Savage T statistic. The performance of this detector is evaluated and compared to that obtained from manually classifying seven recorded utterances with 40, 30, 20, 10, and 0 dB signal-to-noise ratios. In limited testing, the average probability of misclassification is less than 6%, 12% and 46% for signal-to-noise ratios of 39, 20, and 0 dB respectively.

The Constant-Q Transform

Jim Kajiya

A generalization of the short-time Fourier transform is presented which performs constant-percentage bandwidth analysis of time-domain signals. The transform is shown to exhibit frequency-dependent time and frequency resolution. A synthesis transform is also developed which provides an analysis-synthesis system which is an identity in the absence of spectral modification (given a mild analysis window constraint).

SUPPRESSION OF NOISE IN SPEECH USING THE SABER METHOD

Steven F. Boll

ABSTRACT

A stand alone noise suppression algorithm is described for reducing the spectral effects of acoustically added noise in speech. A fundamental result is developed which shows that the spectral magnitude of speech plus noise can be effectively approximated as the sum of magnitudes of speech and noise. Using this simple phase independent additive model, the noise bias present in the short time spectrum is reduced by subtracting off the expected noise spectrum calculated during nonspeech activity. After bias removal, the time waveform is recalculated from the modified magnitude and saved phase. This Spectral Averaging for Bias Estimation and Removal, or SABER method requires only one FFT per time window for analysis and synthesis.

Summary

Background

The majority of narrow-band speech compression algorithms were designed and tested based upon noise-free speech as input. However, the systems constructed from these algorithms will be used in both quiet and noisy environments. For the noise environments such as the helicopter cockpit, the intelligibility and quality of transmitted compressed speech must be maintained at an acceptable level. Methods available to suppress noise in actual operating environments include modifying the speech compression system, use of noise cancelling microphones, or the insertion of a preprocessing noise suppression system prior to vocoder input. This paper describes a preprocessing noise suppression algorithm. This approach was chosen since one, the vocoder system is not modified, two, the noise suppression algorithm is now independent of any specific vocoder implementation, three, most noise cancelling microphones do not generally remove noise above about 1kHZ [1], and four, the method proposed is straightforward to implement and can run in real time. Below are summarized the objectives, approach, and results of this technique.

Objectives

Develop a model for characterizing the spectral effects of additive noise on speech. Insure that the model be applicable simultaneously to both narrow-band periodic noise and wide band colored noise. Minimize the number of apriori assumptions needed to justify the model or implement the algorithm based on the model. Insure that in the absence of noise that the algorithm reduces to essentially an identity system.

Design and implement a noise suppression algorithm based on the model having digital speech in and digital speech out. To afford low cost, real time implementation, keep the implementation as simple as possible, use straightforward estimation techniques and minimize the amount of external information required for effective implementation.

Test the algorithm on speech obtained in realistic operating environments. The speech should be corrupted with noise generated by the environment and acoustically added to the speech. The tests should measure improvements in both intelligibility and quality by comparing results with and without noise suppression.

Tandem the algorithm with a representative narrow-band voice processor. Retest synthetic speech for intelligibility and quality with and without noise suppression preprocessing.

Specify the advantages, limitations and requirements needed for a real time implementation.

The following assumptions were used in implementing the algorithm. The background noise is acoustically or digitally added to the speech. The background noise environment remains locally stationary to the degree that its spectral magnitude expected value just prior to speech activity equals its expected value during speech activity. If the environment changes to a new stationary state, there exists enough time (about 300 ms) to estimate a new background noise spectral magnitude expected value before speech activity commences. For the slowly varying non-stationary noise environment, the algorithm requires a speech activity detector to signal the program that speech has ceased and a new noise bias can be estimated. Finally that significant noise reduction is possible by removing the effect of noise from the magnitude spectrum only.

Basis for Analysis. The fundamental property is developed which demonstrates that the spectral magnitude of noisy speech can be effectively modeled as the sum of magnitudes of speech and noise. As such the additive noise exhibits itself as possibly a wide variance bias added to the desired speech spectrum. Therefore an estimate of the speech magnitude spectrum is obtained by subtracting off an estimate of the noise bias. If the noise has primarily a wide variance non-deterministic component, then local

averaging of magnitude spectra is used to reduce the noise variance. If the noise is primarily narrow variance then no averaging is required for variance reduction prior to bias removal.

Method. Speech is analyzed by windowing data from half-overlapped input data buffers. The magnitude and phase spectra of the windowed data is calculated and the phase is saved. Magnitudes from adjacent windows are then averaged and the spectral noise bias calculated during non-speech activity is subtracted off. Resulting negative amplitudes are then either rectified or zeroed out. A time waveform is recalculated from the modified magnitude and saved phase. This waveform is then overlap added to the previous data to generate the output speech.

Advantages and Limitations. The method requires only a single microphone. It is applicable to both wide-band and narrow-band noise sources. The method is computationally efficient requiring only one FFT per analysis frame with the FFT computation per frame increasing logarithmically with the sampling rate. Finally, the algorithm output is speech and thus can be tandemed to any narrow-band speech processor.

Limitations of the algorithm include the requirement of a locally stationary noise environment and possibly a speech activity detector for updating the noise bias estimate following a spectral noise shift. If the noise is non-coherent, then the averaging required for variance reduction will produce some temporal echo-like smearing. In addition as will be shown, the spectral estimation error is proportional to the amount of variance reduction. Therefore, only partial noise cancellation is possible for wide variance noise sources.

Results. The performance of the SABER algorithm has been initially measured using a limited Diagnostic Rhyme Test (DRT). Testing was conducted by Dynastat, Inc. [2] using clear channel and helicopter noise tapes. Measures for improvements in intelligibility as well as a course measure of quality were conducted using a single speaker test. Results indicate average improvements in intelligibility with some subareas having major improvements and major improvements in quality. Detailed scores are given below in section on results.

Input-Output Data Manipulation

Speech from the A-D converter is segmented and windowed such that in the absence of spectral modifications when the synthesis speech segments are added together, the resulting overall system reduces to an identity. The data is segmented and windowed using on the result [3] that if a sequence is separated into half-overlapped data buffers, and each buffer is multiplied by a Hanning window, then the sum of these windowed sequences add back up to the original sequences. The window length is chosen to be approximately twice as large as the maximum expected pitch period for adequate frequency resolution [4]. For the sampling rate of 8.00 kHz a window length of 256 points shifted in steps of 128 points was used. Figure 1 shows the data segmentation and advance:

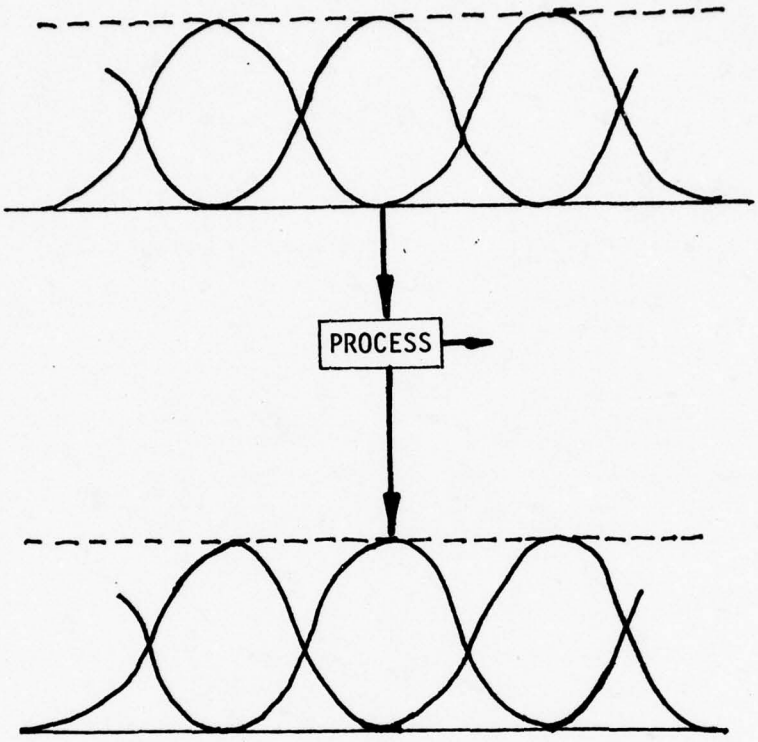


FIGURE 1 DATA SEGMENTATION

Frequency Analysis

The DFT of each data window is taken and converted to the polar coordinates of magnitude and phase.

Since real data is being transformed, two data windows can be transformed using one FFT [5]. The FFT size is set equal to the window size of 256. Augmentation with zeros was not incorporated. As correctly noted by J. Allen [6], spectral modification followed by inverse transforming can distort the time wave-form due to temporal aliasing caused by circular convolution with the time response of the modification. Augmenting the input time waveform with zeros before spectral modification will minimize this aliasing. Experiments with and without augmentation using the helicopter speech resulted in negligible differences and therefore augmentation was not incorporated. Finally, since real data is analyzed transform symmetries were taken advantage of to reduce storage requirements essentially in half.

Magnitude Averaging

As is shown below, the variance of the noise spectral estimate is reduced by averaging over as many spectral magnitude sets as possible. However, the non-stationarity of the speech limits the total time interval available for local averaging. The number of averages is limited by the

number of analysis windows which can be fit into the stationary speech time interval. The choice of window length and averaging interval must compromise between conflicting requirements. For acceptable spectral resolution a window length greater than twice the expected largest pitch period is required with a 256 point window being used. For minimum noise variance a large number of windows are required for averaging. Finally, for acceptable time resolution a narrow analysis interval is required. A reasonable compromise between variance reduction and time resolution appears to be three averages. This results in an effective analysis time interval of 38 ms.

Bias Estimation

The SABER method requires an estimate at each frequency bin of the expected value of noise magnitude spectrum, μ_N:

$$\mu_N = E\{|N|\}$$

This estimate is obtained by averaging the signal magnitude spectrum $|X|$ during non speech activity. Estimating μ_N in this manner places certain constraints when implementing the method. If the noise remains stationary during the subsequent speech activity, then an initial startup or calibration period of noise-only signal is required. During this period (on the order of a third of a second) an estimate of μ_N can be computed. If the noise environment

is nonstationary then a new estimate of μ_N must be calculated prior to basis removal each time the noise spectrum changes. Since the estimate is computed using the noise-only signal during non-speech activity, a voice switch is required. When the voice switch is off an average noise spectrum can be recomputed. If the noise magnitude spectrum is changing faster than an estimate of it can be computed, then time averaging to estimate μ_N cannot be used. Likewise if the expected value of the noise spectrum changes after an estimate of it has been computed, then noise reduction through bias removal will be less effective or even harmful.

Bias Removal

The SABER spectral estimate $\bar{s}_A$ is obtained by subtracting the expected noise magnitude spectrum μ_N from the averaged magnitude signal spectrum $\overline{|X|}$

Thus:

$$\bar{s}_A(k) = \overline{|X(k)|} - \mu_N(k) \quad k = 0, 1, \ldots, L-1$$

where L =DFT buffer length.

After subtracting, the differenced values having negative magnitudes can either be set to zero (half rectification) or be made positive (full rectification). These negative differences represent frequencies where the sum of speech plus local noise is less than the expected

noise. As referenced below, full-wave rectification minimizes the spectral error. However, if the noise source drops out during speech, full rectification will result in the expected noise value being incorrectly added back in to the speech spectrum. This in fact happened for the helicopter tapes processed. Therefore half rectification was used. Figures 2 and 3 show input-output frequency relations for half and full rectification.

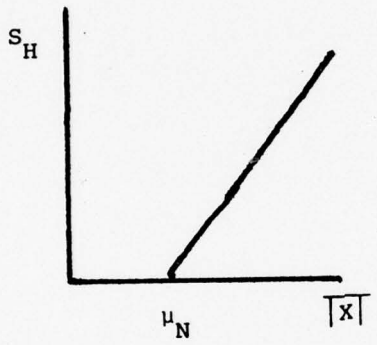


Figure 2

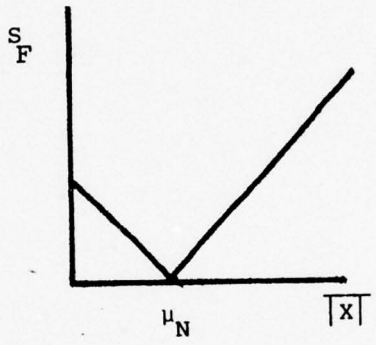


Figure 3

Input - Output Relations

Synthesis

After bias removal and rectification, a time waveform is reconstructed from the modified magnitude and the phase buffer corresponding to the center window. Again since only real data is generated, two time data sets are computed simultaneously using one inverse FFT. The data windows are

then overlap added to form the output speech sequence. The overall block diagram is shown in Figure 4.

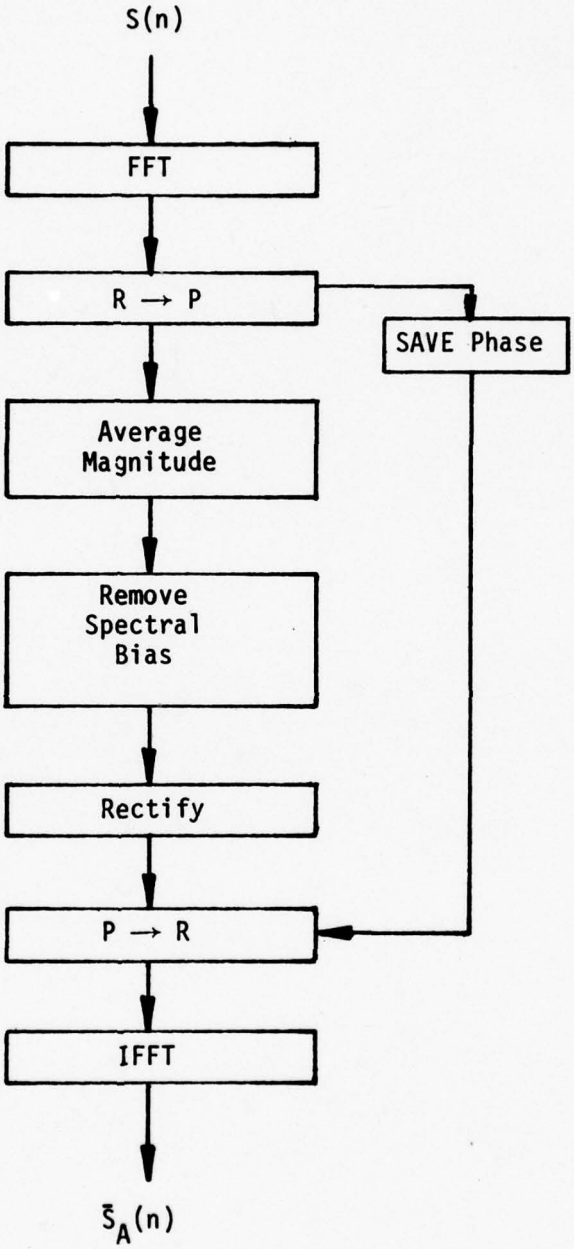


Figure 4
Block Diagram

Results

The ability of this method to improve intelligibility is being measured using the Diagnostic Rhyme Test (DRT) [2]. A measure of quality improvement is also available using the DRT data base [7]. This section lists preliminary results for a limited DRT test using a single speaker. The data, provided by Dynastat, Inc., consisted of speaker RH recorded in a helicopter environment. The results are given using Tables 1 and 2. Table 1 list intelligibility scores for the original data and the SABER output, followed by intelligibility scores for an LPC vocoder output which used original or SABER as input. Table 2 list quality scores of original and SABER followed by quality scores of LPC output using either original or SABER as input.

	Original	SABER	LPC on Original	LPC on SABER
Voicing	95	91	84	86
Nasality	82	77	56	52
Sustension	92	86	49	56
Sibilation	75	84	61	88
Graveness	68	66	61	59
Compactness	88	88	83	93
Total	84	82	66	72

Table 1

DRT Scores for Single Speaker

	Original	SABER	LPC on Original	LPC on SABER
Naturalness	49	47	40	41
Inconspicuousness of Background	30	41	29	38
Intelligibility	31	30	22	26
Pleasantness	16	26	13	23
Acceptability	28	32	22	28
Total	25	29	19	25

Table 2

Quality Scores from DRT Data

Observations

This single speaker DRT test indicates that SABER processing followed by LPC significantly increases intelligibility. Scores in the areas of voicing, nasality and graveness are about equal. It improves the apprehensibility of sustension, sibilation, and compactness.

The quality measures taken clearly indicate that SABER enhances listener acceptability. The background noise is less conspicuous, and the processed speech more pleasant.

Analysis of the Phase Independent Model

Assume that a noise signal n, has been added to a speech signal s, with their sum denoted as x.

Then

$$x = s + n$$

Taking the Fourier transform gives

$$X = S + N$$

The desired speech spectral magnitude, $|S|$ is given by

$$|S| = |X - N|$$

The zero phase approximation S_z to $|S|$ is given by

$$S_z = |X| - |N|$$

When S_z goes negative it can be half-rectified S_H or full-rectified, S_F:

$$S_H = \frac{S_z + |S_z|}{2}$$

$$S_F = |S_z|$$

The spectral error D at any frequency is given by

$$D_H = |S| - S_H = |S| - \left(\frac{S_z + |S_z|}{2}\right)$$

$$D_F = |S| - S_F = |S| - |S_z|$$

It can be shown [8] that the full-rectified modeling error is zero for $|N|>|S|$ and the relative error $D/|S|$ inversely proportional to the signal to noise ratio for $|S|>|N|$. For $|X|>|N|$ the half-rectified modeling error will increase to as much as $|S|$. However, if the noise floor were to suddenly decrease well below its average value, the full-rectified estimate would incorrectly add noise into the estimate whereas the half-rectified estimate would not. Thus the half-rectified estimate would give better results in this situation.

Analysis and Reduction of Estimation Error Error Estimate

Using the zero phase model (assuming $|X|>|N|$ for simplicity) the SABER estimation error is given by

$$\epsilon = S_A - S_z = |X| - \mu_N - |X| + |N|$$

where

$S_A = |X| - \mu_N$ equals unaveraged SABER estimate

$\mu_N = E\{|N|\}$ equals expected noise magnitude spectrum

$S_z = |X| - |N|$ zero phase estimate of $|S|$

Thus the spectral error ε equals, $|N| - \mu_N$, the difference between the magnitude of the noise spectrum and its expected value.

Averaging

The spectral error can be reduced by averaging magnitude spectral $\overline{|X|}$. The amount of reduction by averaging has been carefully investigated [8], [9]. For example, if five half-overlapped windows are used [8]:

$$E\{(\overline{|N|} - \mu_N)^2\} = 0.275 \text{ var } \{|N|\} = (0.06)\sigma_N^2 L$$

This gives a total variance reduction of -12.4 dB.

References

1. Dave Coulter, Private Communication.

2. William D. Voiers, Alan D. Sharpley, and Carl H. Hehmsoth, Research on Diagnostic Evaluation of Speech Intelligibility, ,Final Report AFSC Contract No. F19628-70-C-0182 1973

3. T.W.Parsons and M. R. Weiss, Enhancing Intelligibility of Speech in Noisy Environments or Multi-Talker Environments, Final Report RADC-TR-75-155 Contract No F30602-74-C-0175, 1975

4. John Makhoul and Jerry Wolf, Linear Prediction and the Spectral Analysis of Speech, NTIS No. AD-749066, BBN Report No. 2304, Bolt Beranik and Newman Inc. 1972.

5. O. Brigham The Fast Fourier Transform, Englewood Cliffs, New Jersey, Prentice Hall, 1974.

6. J. Allen, "Short Term Spectral Analysis, Synthesis, and Modification by Discrete Fourier Transform," IEEE Trans. on Acoust., Speech and Signal Proc., Vol. ASSP-25, No.3, June 1977.

7. In house research, Dynastat Inc. Austin Texas, 78705.

8. Steven F. Boll, Noise Suppression Methods for Robust Speech Processing, Semi- Annual Technical Report UTEC-CSc-77-202, Contract No. N00173-77-C-0041, 1977.

9. Peter Welch, "The Use of the Fast Fourier Transform for the Estimation of Power Spectra: A Method Based on Time Averaging Over Short, Modified Periodograms," IEEE Trans. Audio Electroacoust., Vol. AU-15, June 1967.

A SUMMARY OF RECENT EXPERIMENTS APPLYING ADAPTIVE NOISE CANCELLATION TECHNIQUES TO AUDIO SIGNALS

Dennis Pulsipher

Introduction

A dual input noise cancellation technique for audio signals was presented in a semi-annual report a year ago. The philosophy behind the technique was quite different from that of traditional techniques. Instead of estimating the desired signal directly, the technique attempted to estimate the noise directly and obtained a signal estimate by subtracting the noise estimate from the noisy signal.

The experiments which had been performed at that time used synthetic data and demonstrated great potential for the technique. In the last semi-annual report initial experiments in a real environment were described. A description of experiments that have followed and some of the questions they have raised comprises the body of this report.

The Experiments

Upon successful completion of the synthetic results described in previous reports, experiments to evaluate real situations were begun. An attempt was made to design these experiments so that real acoustic situations were used, without completely destroying the validity of the assumed data generation model.

Efforts to maintain a certain amount of consistency between experiments resulted in a set of control conditions which were maintained constant for all recent experiments. To minimize recording effects, it was decided to digitally record the noisy signal and the noise reference signal simultaneously. All signals were low-pass filtered to a bandwidth of 3.2 kHz and sampled at a rate of 6.67 kHz. Control of the environment was maintained by recording in a single, isolated, but acoustically live room. While no effort was made to simulate a point noise source, the noise was generated at the side of the room by a single, high quality speaker system, which was kept in a fixed position.

The initial experiments performed used two microphones separated by approximately 8 feet, located near the middle of the room, to pickup the noisy

channel and the noise reference channel. By using slow adaption rates (time constants of approximately 5 seconds) and long transversal filter lengths (3000 points), noise reduction of approximately 16 dB was achieved. Doubling the length of the filter resulted in about 1 dB improvement over that level.

Since synthetic experiments had yielded significantly better results, questions were raised about the validity of treating a room as a linear channel, whether or not small movements in the room affected stationarity assumptions, if the lack of a point noise source was a serious complication, or if something about the channels themselves was the cause of the degradation.

To identify the sources of degradation another series of experiments was undertaken. Empirical estimates of the impulse response of the room from a single source to two separate points in the room were made. A known digitized noise source was then digitally filtered through the two different impulse responses measured. These filtered noise samples were then used as the noisy signal and reference noise inputs to the noise cancellation algorithm. Thus, the acoustically recorded experiments were simulated with similar channels to to those expected, but wherein linearity and stationarity were guaranteed. These

experiments yielded results roughly 2 dB better than the corresponding experiments which used acoustically produced data. Differences in microphone placement and lack of additional uncorrelated noise at low levels were considered capable of causing such minor differences, and it was concluded that assumptions of both stationarity and linearity of the channels were probably justified. It was also concluded that the lack of a point source was not a serious problem.

At this point it was strongly suspected that the fact that one of the channels had to be effectively inverted was the cause of the degradation. Many theoretical and practical issues regarding inverse filtering were considered and it was decided to devise a quick experiment to verify this suspicion.

Since the major obstacle to great success with acoustic data appeared to be involved with inverting one of the room's channels, it was decided to see if the problem could be eliminated by forcing that channel to be an identity system, which could be trivially inverted. The noisy signal, was therefore recorded as before, with a microphone placed in the middle of the room. The reference noise, however, was not recorded through a microphone at all, but directly from the electrical signal used to drive the speaker system. This configuration achieved noise reduction of

approximately 26 dB confirming suspicions that effective channel inversion was a major problem.

It was then wondered if careful placement of the noise reference pick-up microphone might be used to improve results by making the channel needing inversion appear to be a near identity system. The acoustic experiment was repeated with the noisy signal being recorded from the middle of the room, and the noise reference being recorded by a microphone placed directly facing the high energy output section of the speaker system. Results comparable to the simulated room experiments were obtained from this experiment (18 dB noise reduction). While this technique may be effective if a real-time system is available to search for an optimal position for reference noise collection, results indicated that it was not simply a matter of closeness of microphone placement to the noise source which was going to be a final solution.

Conclusions

During these experiments it became obvious that many facets of the noise cancellation problem are yet to be understood. Techniques dealing with filter inversion are being investigated to better understand the problems involved. Investigations are also underway to improve convergence of channel estimates when frequency bands of low energy are contained in the reference noise samples. Even if these investigations are fruitless, however, noise cancellation now appears to be a worthwhile approach to signal restoration in acoustically hostile environments.

Estimation of the Parameters of an Autoregressive Moving-Average Process In the Presence of Noise

William J. Done

The previous report on this project presented the details for the autoregressive moving-average (ARMA) process generated by adding white noise to an autoregressive (AR) process. That report stressed the problems that are inherent in estimating the parameters of the resulting ARMA process. Part of this estimation problem lies in the validity of this model for a given application. A major part of the difficulty, however, lies in developing estimation procedures for ARMA processes, regardless of the source of that process. The primary effort in this project since the last report has been the investigation of various methods that might be used to estimate the autoregressive and moving-average coefficients of an ARMA process from data generated by that process. Three methods have been implemented for evaluation.

One procedure mentioned in the previous report is the Mode 1 iterative method by Steiglitz and McBride [4]. The approach is: given input and output sequences for an unknown system, determine the filter which approximates the unknown system by using a filter which is the ratio of two rational polynomials (in the Z-domain). Graphically, the problem is illustrated in Figure 1. The polynomials A(z) and B(z) are given by

$$A(z) = \sum_{i=o}^{q} a(i)z^{-i}, \quad a(o) = 1$$

and

$$B(z) = \sum_{i=0}^{p} b(i)z^{-i}$$

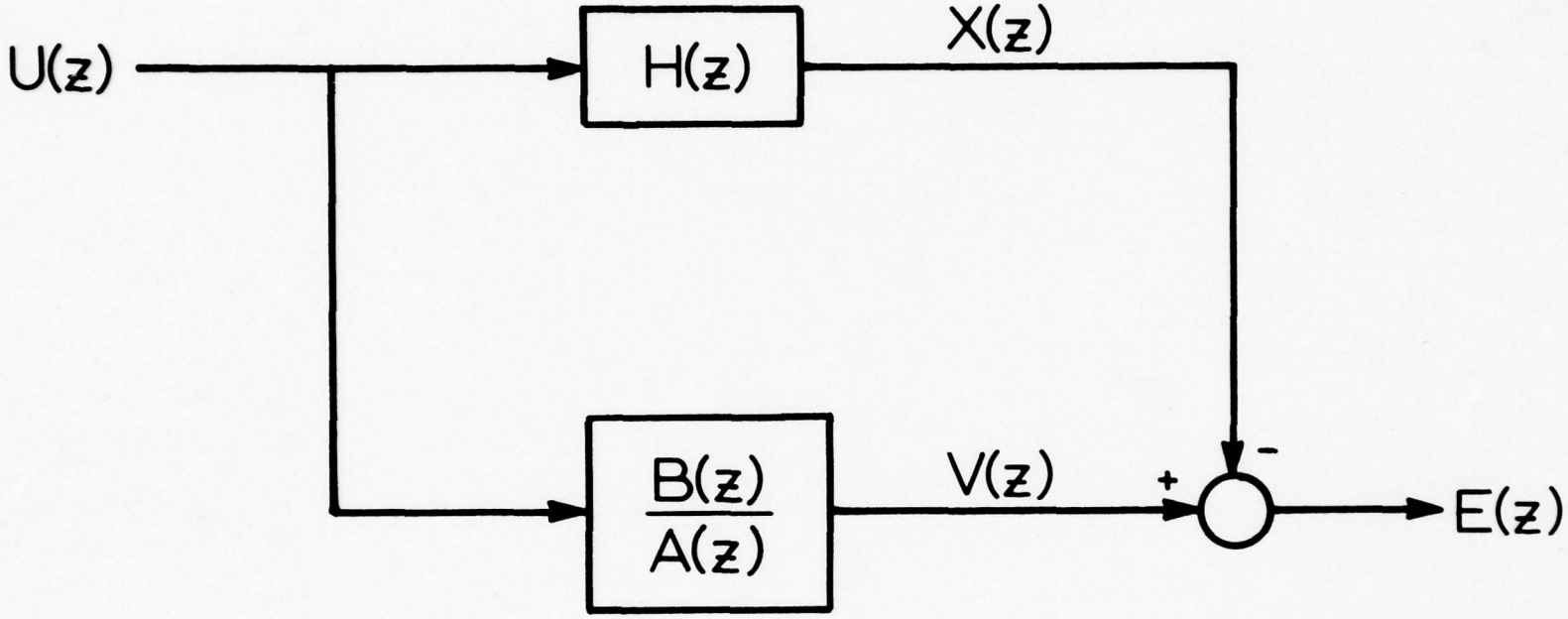


Figure 1: Block Diagram of System Identification Model Used in Mode 1 Iterative Procedure of Estimating Parameters.

The coefficients a(i) and b(i) in A(z) and B(z), respectively, are selected to minimize E(z) in some sense. The model's response, V(z), is

$$V(z) = \frac{B(z)}{A(z)} U(z) \qquad 1)$$

or

$$A(z)\ V(z) - B(z)\ U(z) \qquad 2)$$

Also, from the block diagram, we have

$$E(z) = V(z) - X(z) \qquad 3)$$

Steiglitz then performs a "quasi-linearization" on 2), using previous iterations to form approximations to the derivatives,

$$A^{i}(z)V^{i}(z) + [A^{i+1}(z) - A^{i}(z)]V^{i}(z) + A^{i}(z)[V^{i+1}(z) - V^{i}(z)] = B^{i+1}(z)\ U(z) \qquad 4)$$

where the superscript indicates the iteration number. Replacing V (z) with X(z) in 4) and simplifying gives

$$A^{i}(z)V^{i+1}(z) = [A^{i}(z) - A^{i+1}(z)]\ x(z) + B^{i+1}(z)\ U(z) \qquad 5)$$

Solving for V (z) and using that expression for V(z) in 3) gives

$$E^{i+1}(z) = V^{i+1}(z) - X(z) = \frac{B^{i+1}(z)}{A^{i}(z)} U(z) - \frac{A^{i+1}(z)}{A^{i}(z)} X(z) \qquad 6)$$

It is the form of 6) that suggests the Mode 1 technique presented by Steiglitz. Noting that both U(z) and X(z) are

recursively filtered through the i^{th} iteration of A(z), define $\hat{U}(z) = U(z)/A(z)$ and $\hat{X}(z) = X(z)/A(z)$. With these definitions, the time domain representation for 6) is

$$e(k) = \sum_{i=0}^{p} b(i)\,\hat{u}(k-i) - \sum_{j=o}^{q} \hat{x}(j)a(k-j) \qquad 7)$$

where the iteration number has been dropped for convenience. The coefficients $\{a(i)\}_1^q$ and $\{b(i)\}_1^p$ are selected to minimize e(k) in the least squares sense. The least squares procedure requires the solution of the matrix equation

$$\underline{R}_{ux}\,\underline{\delta} = \underline{r}_{ux} \qquad 8)$$

where $\underline{R}_{ux}$ is a matrix composed of the auto- and cross-correlations of u(k) and x(k); $\underline{r}_{ux}$ is a vector composed of those correlations; and δ is the solution vector containing the desired a(i) and b(i) coefficients. Use of this method thus requires the solution of a set of $p + q + 1$ linear simultaneous equations.

For application to the estimation of the coefficients of an ARMA process, this technique must be modified slightly. When only the output of the system is known, u(k) is assumed to be the Kronecker delta function. Also, the system output x(k) may be modified so that it more closely resembles an impulse response, as the assumption for u(k) implies. To test Steiglitz's Mode l method, the following tests have been performed:

1) From a known model, considered to be the system to be identified, generate an output sequence x(k).

2) The input to the "unknown" system is one of: impulse, impulse train, or noise (approximately white).

3) Use the Mode 1 method to compute estimates for the parameters of the "unknown" system.

4) Compare the parameter estimates to the design parameters.

The results for one 10-pole, 2-zero model system are now presented. This model is identical to that used by Steiglitz in [3].

Figure 2 is the model spectrum to be identified. Note that the zeros are a complex conjugate pair located on the unit circle. Figure 3 is the output of this model when excited by an impulse. Figure 3a) is the time sequence and Fig. 3b) is the spectrum of that sequence in dB. The estimate for the model spectrum produced by one iteration of this method is shown in Fig. 4b). The model spectrum is repeated in Fig. 4a).

The time sequence produced by exciting the model with an impulse train is shown in Fig. 5a). The period for this example is 100. In Fig. 5b) is the estimate of the spectrum of this process obtained by Hamming windowing the time sequence and performing a DFT. The result is in dB. In using the Mode 1 technique for this type of time sequence, it is desirable to preprocess x(k) to make it more like an impulse response. Figure 6a) shows the real part of the complex cepstrum of x(k) after Hamming windowing x(k).

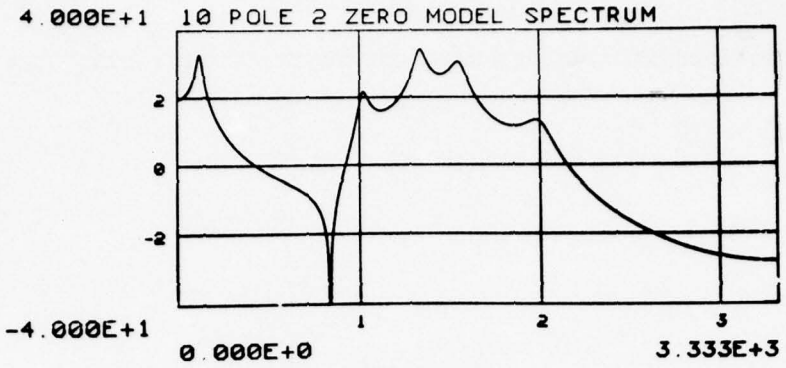


Figure 2: 10 Pole, 2 Zero Model Spectrum to be Identified.

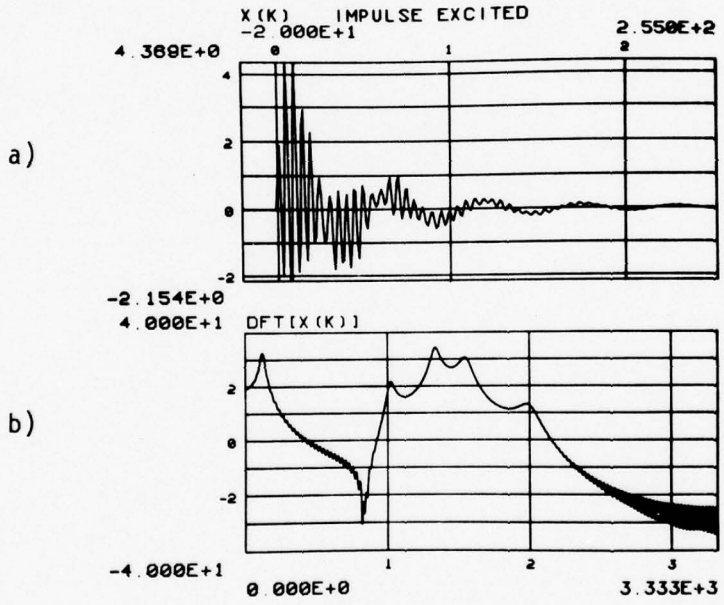


Figure 3: a) Output of Model with an Impulse Input.
b) Spectrum of Part a).

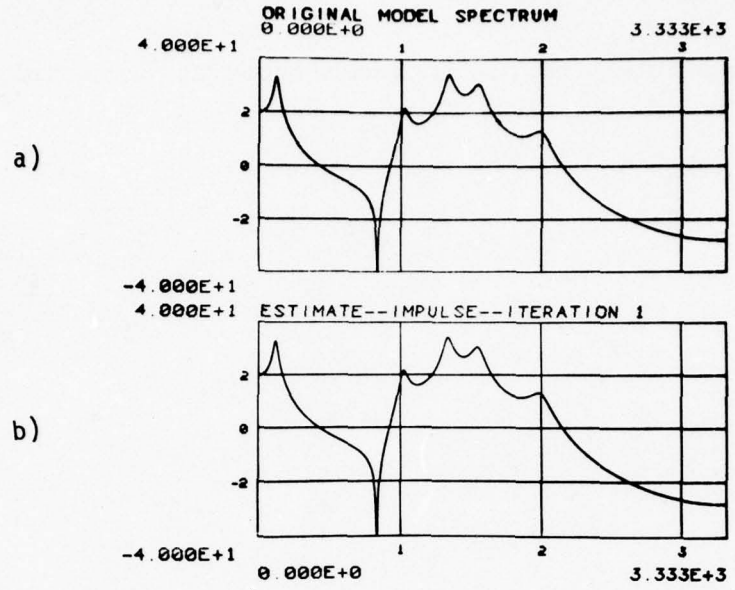


Figure 4: a) Model Spectrum.
b) Estimate of Model Spectrum after 1 Iteration, Impulse Excitation.

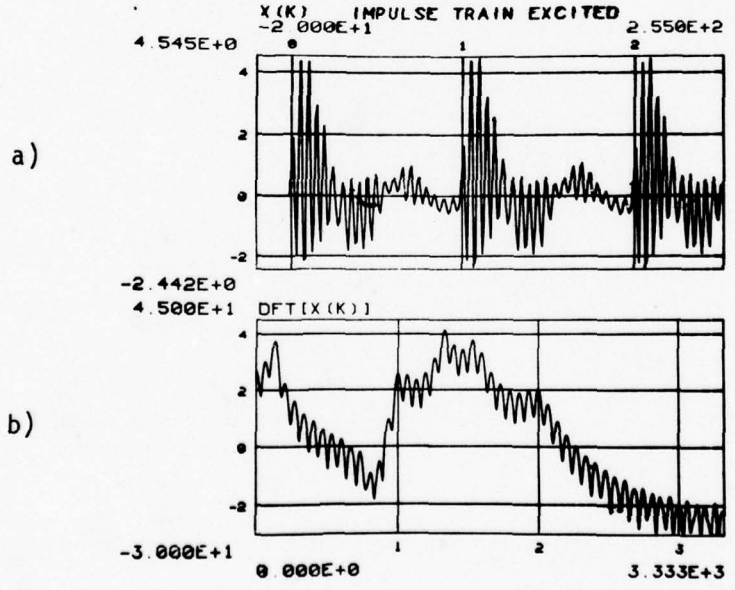


Figure 5: a) Output of Model with an Impulse Train Input.
b) Spectrum of Part a).

By properly windowing this cepstrum, two useful tasks are accomplished. First, by eliminating the spikes resulting from the harmonics in the frequency domain, the periodic nature of x(k) can be suppressed. The second step is to force this cepstral representation of x(k) to be causal. Upon returning to the time domain, if appropriate scaling has been done in the cepstrum, the resulting time series will be minimum phase. Figure 6b) shows the cepstrum after windowing. Figure 7b) contains the new minimum phase time sequence, while Fig. 7a) contains the output of the impulse excited model for comparison. Figures 8a) and 8b) are respectively the spectral estimates of x(k) and $x_{mp}(k)$, the modified version of x(k). Note in Fig. 7 that $x_{mp}(k)$ is quite similar to x(k) from the impulse excited case. Figure 8 shows the suppression of the harmonic structure on the spectrum of x(k) caused by the periodic nature of x(k). The Mode 1 technique is now applied to $x_{mp}(k)$. Figure 9b) shows the estimated spectrum for the impulse train excited case after two iterations. The original model spectrum is repeated in Fig. 9a).

The last case to be considered is when the model has been excited by a noise sequence. The resulting output sequence and spectral estimate are shown in Fig. 10a) and 10b), respectively. Superimposed on the spectrum of the noise excited x(k) is the original model spectrum. Note the random variations from that ideal spectrum resulting from the deviation of the excitation sequence from an ideal white

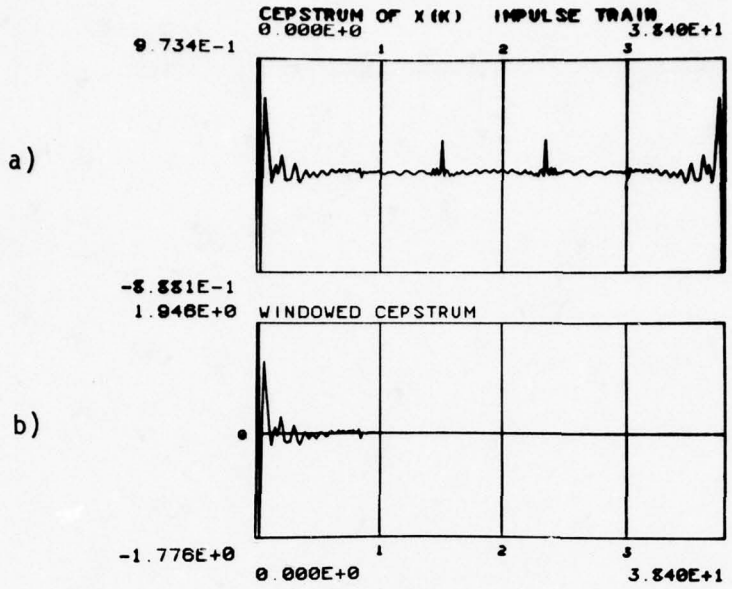


Figure 6: a) Real Part of the Complex Cepstrum of x(k), (Impulse Train Excited).
b) Part a) after Cepstral Windowing.

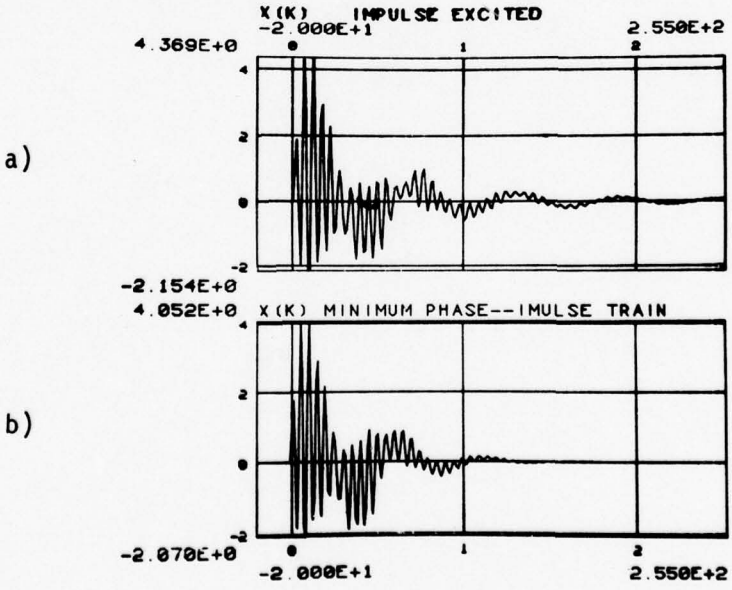
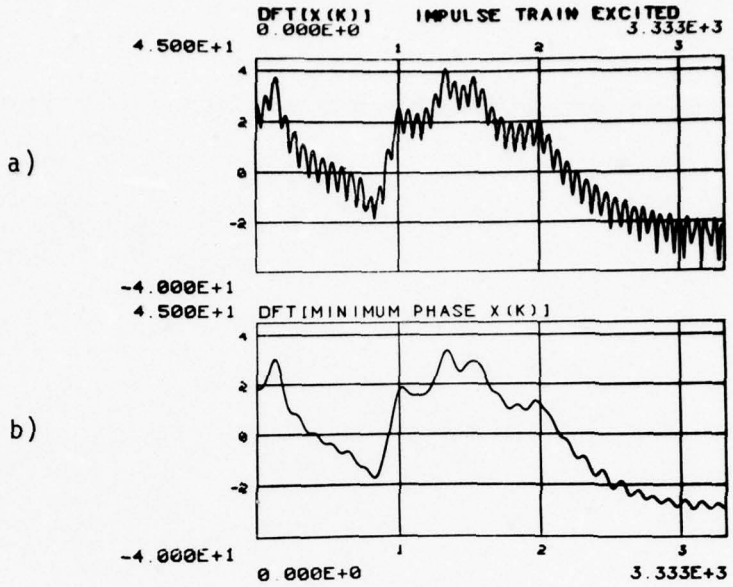


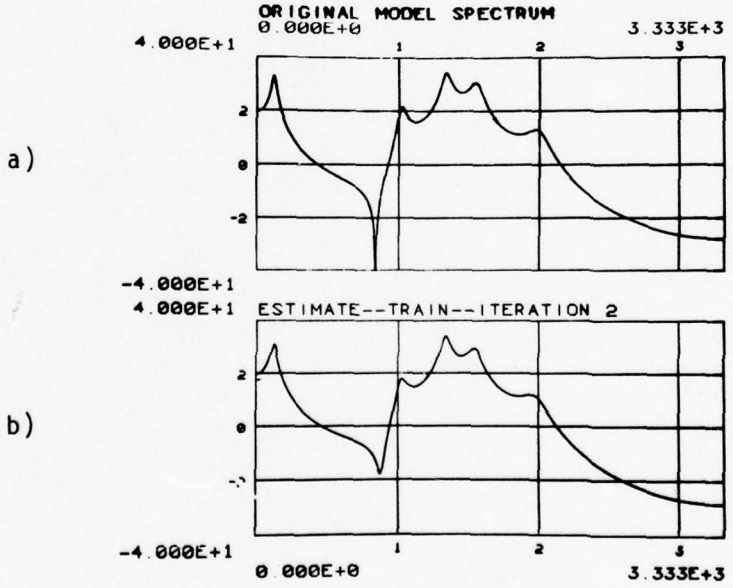
Figure 7: a) Output of Model with an Impulse Input.
b) Modified System Output after Cepstral Modifications to Remove Periodicity and Simulate a Minimum Phase Signal.



a)

b)

Figure 8: a) Spectrum of x(k) Produced by an Impulse Train Input to the Model.
b) Spectrum of the Cepstrally Modified Version of x(k).



a)

b)

Figure 9: a) Model Spectrum
b) Estimate of Model Spectrum after 2 Iterations, Impulse Train Excitation.

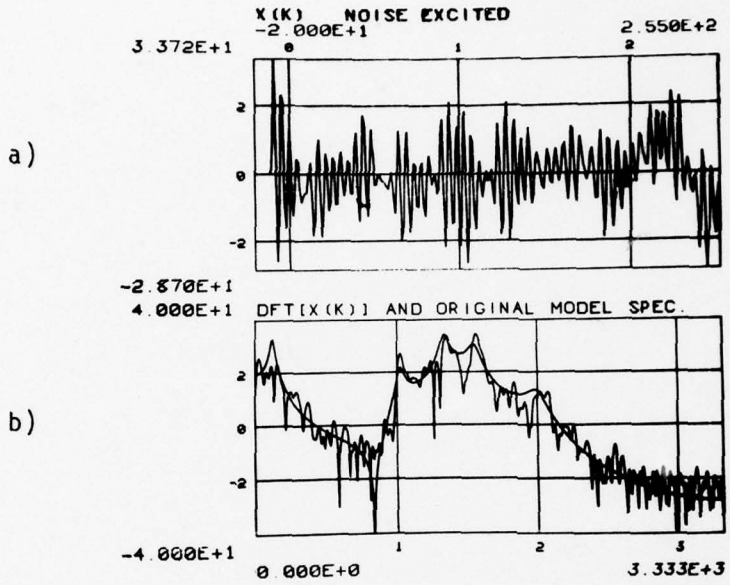


Figure 10: a) Output of Model with Noise as the Input.
b) Spectrum of Part a).

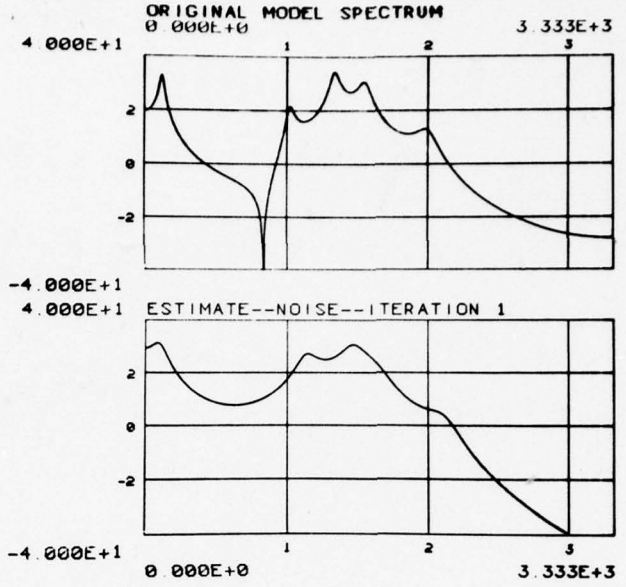
noise process. Figures 11b) and 12b), respectively, show the spectral estimates produced by this technique after the first and second iterations. Further iterations fail to improve upon the the estimate, which is poor.

Because of the results for the noise excited case, this method has been discarded. The estimates obtained for noise excited processes were consistently poor, often converging to unstable filter estimates. In addition, the Mode 1 method is strongly dependent upon double precision arithmetic to achieve success, even in the impulse and impulse train excited cases.

The parameter estimation presently being investigated is that given in Anderson [1]. This method is based on a time domain Newton-Raphson approach to maximization of the log likelihood function. Anderson's approach in this method is to assume zero initial conditions for the data $x(k)$, $k < 0$. If $Q(\theta)$ is the log likelihood function to be maximized, where θ is the vector of coefficients $\{a(i)\}_1^q$ and $\{b(i)\}_1^q$ to be estimated, then

$$g_i(\theta) = \frac{\partial Q(\theta)}{\partial \theta_i} = 0 \qquad 9)$$

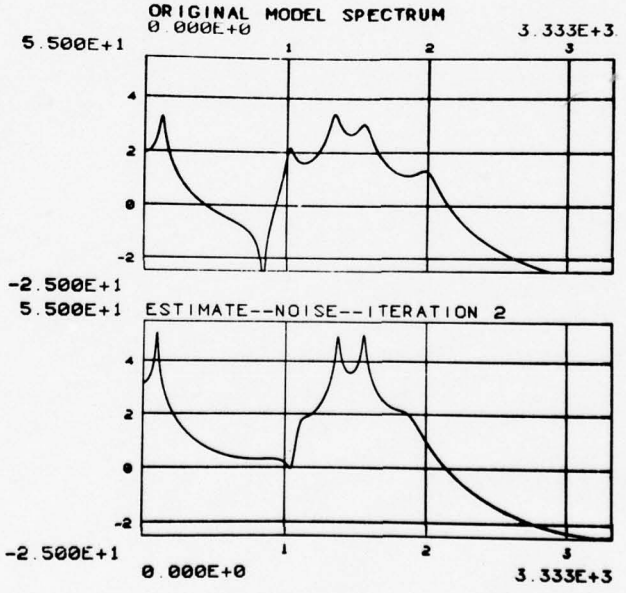
defines a set of $p + q$ equations which must be satisfied by appropriate choice of θ. If the log likelihood function results from the assumption of normality for the excitation noise, then the solution of 9) will be the maximum of $Q(\theta)$. The relationships in 9) are nonlinear, however, so a Taylor



a)

b)

Figure 11: a) Model Spectrum.
b) Estimate of Model Spectrum after 1 Iteration, Noise Excitation.



a)

b)

Figure 12: a) Model Spectrum.
b) Estimate of Model Spectrum after 2 Iterations, Noise Excitation.

expansion of 9) about the optimal solution θ is performed:

$$g(\theta^*) + g'(\theta^*)\ (\theta-\theta^*) = 0 \qquad 10)$$

Solving 10) for θ, we have the Newton-Raphson iterative method for maximizing the log likelihood function.

The Newton-Raphson method has the advantage of the most rapid convergence when in the neighborhood of the solution if the log likelihood function is approximately quadratic in that region. The method does require the evaluation of the second derivatives of the log likelihood function. Also, if the initial guess for the parameters is not in the neighborhood of the maximum, the convergence rate may be slow or convergence to an incorrect solution may occur. The Newton-Raphson (N-R) method requires the inversion of the Hessian of $Q(\theta)$, which is not guaranteed to be positive semidefinite at θ. As a result, $Q(\theta^{i+1})$ may be less than $Q(\theta^i)$, where the superscript indicates the iteration of θ. There are methods which avoid some of these problems. These will be investigated at a later date.

The N-R estimation technique has been implemented and is presently being evaluated. Its operation has been tested on impulse and noise excited models. For the impulse excited models, the estimate is excellent. In the noise excited sequences, errors do occur in the parameter estimates. Those for strictly AR or MA processes seem to be consistent with the findings of others. The estimates for ARMA processes tend to exhibit more error. In general, the

N-R method seems to provide better estimates of the parameters than Steiglitz's Mode 1 method (especially for the noise excited data), but a high variability in the accuracy of the estimates occurs from frame to frame. Double precision arithmetic does not seem to be necessary for this method.

In an attempt to verify the correct operation of the N-P method, a third estimation procedure has been implemented. This is a direct search method based on the unconditional sum of squares method discussed in Box and Jenkins [2]. This method is inefficient computationally because the sum of squares of the estimated excitation sequence must be computed at a large number of points in the parameter space. Its usefulness lies in allowing one to view the shape of the space for low order cases. For the AR(1) process degraded by white noise, the observed data is an ARMA(1,1) process. This ARMA(1,1) process is generated by computing the moving-average coefficient and excitation variance that would result from adding white noise to the AR(1) process. The ARMA(1,1) process is then generated directly, as opposed to generating the ARMA(1,1) process by creating the AR(1) process and adding the white noise. Using both the N-R method and direct search method, tests performed on this first order model have been useful in predicting how this approach will respond for varying levels of noise. Further tests on the N-R method and modifications are now being conducted.

REFERENCES

1. T.W. Anderson, "Estimation for Autoregressive Moving Average Models in the Time and Frequency Domains," The Annals of Statistics, vol. 5, no. 5, 1977, pp. 842-865.

2. G. E. Box and G. M. Jenkins, Time Series Analysis, Forecasting and Control, Holden-day, San Francisco, 1976, pp. 208-231.

3. K. Steiglitz, "On the Simultaneous Estimation of Poles and Zeros in Speech Analysis," IEEE Trans. on Acoustics, Speech, and Signal Processing, vol. ASSP-25, no. 3, June 1977, pp. 229-234.

4. K. Steiglitz and L. E. McBride, "A Technique for the Identification of Linear Systems," IEEE Trans. on Information Theory, vol. IT-21, no. 4, July 1975, pp. 476-480.

NONPARAMETRIC-RANK ORDER STATISTICS APPLICATIONS

TO ROBUST SPEECH ACTIVITY DETECTION

Benjamin V. Cox

Abstract

This report describes a theoretical and experimental investigation for detecting the presence of speech in wideband noise. An algorithm for making the silence-speech decision is described. This algorithm is based on a nonparametric statistical signal-detection scheme that does not require a training set of data and maintains a constant false alarm rate for a broad class of noise inputs. The nonparametric decision procedure is the multiple use of the two-sample Savage T statistic. The performance of this detector is evaluated and compared to that obtained from manually classifying seven recorded utterances with 40, 30, 20, 10, and 0 dB signal-to-noise ratios. In limited testing, the average probability of misclassification is less than 6%, 12% and 46% for signal-to-noise ratios of 39, 20, and 0 dB respectively.

Introduction

The problem of detecting voice signals in the presence of noise has only been addressed by a small number of investigations. In these investigations, the traditional approach to distinguishing between voice and noise or estimate the bandwidth occupied by the speech signal was to level detect waveform power (1, 2, 3, 4, 10, 24). The threshold normally was experimentally determined by a limited training set of data (1, 4, 5, 6), by the maximum live noise power recommended by CCITT for telephone channels (1, 2, 3, 4, 7, 8, 13) or by a threshold adjustment process updated on a fixed schedule (every half second)(25).

Recently, Atal and Rabiner (21) suggested a pattern recognition approach to voiced-unvoiced-silence classification in which five measurements or features - - energy, zero-crossing rate, autocorrelation coefficient at unit sample delay, first predictor coefficient and energy of the predictor errors were combined using a non-Euclidian distance metric to give a reliable decision. This method was optimized for telephone line inputs by Rabiner, et al, (22) and used for digit recognition by Rabiner, et al (18, 19). The algorithm was modified to do an average signal spectrum template match using the LPC distance measured (23).

L. S. Siegel (67) proposed a modification to the Atal (21) algorithm in which a relatively small set of samples is used to "train" the classifier.

Lin (71), Adoul (28), and Adoul (29) modified Atal and Rabiner's

pattern recognition approach for their proposed detectors.

Reliable discrimination between silence, unvoiced speech, and voiced speech is a difficult problem because no general theory exists which can preselect the optimal features for input to the classifier. Furthermore, robustness is not achieved because parametric statistics that rely on normality assumption were used in almost all the past investigations. The past algorithms also required a training set of data to determine the required detector threshold levels.

In this report, a nonparametric statistical detection technique is used to classify a given interval of speech data as silence or speech and presents results on a limited experiment of seven utterances for various signal-to-noise ratios. The advantages of this technique are that the proposed detection algorithm:

- maintains a constant false-alarm rate (CFAR) at the detector output for large classes of noise distributions
- it is robust (insensitive to changes not under test) and powerful (sensitive to specific factors under test)
- does not require statistical information about either the signal or the background noise (does not require a training set of data)
- performance for signals in non-Gaussian noise may often surpass that of detectors optimized against Gaussian noise

- will operate where the noise statistics are non-stationary or change from one application to another

- simple to implement digitally

- for large n, it is efficient as the Nymann-Pearson detector for a wide class of noise distributions

Nonparametric decision procedures have been previously applied on radar or Sonar systems that have to operate in an environment of heavy external interference (48). The major reason behind the use of this type of detector is its ability to maintain a constant false-alarm rate (CFAR) for large classes of noise distributions (noise, weather, clutter, interference). These detectors can be designed to operate in an environment where very little statistical information about either the signal or the background noise is available. In addition, the detection performance of such detectors for signals in non-Guassian noise may often surpass that of detectors optimized against Gaussian noise (48, 60, 63).

This study has been primarily concerned with detectors based on rank order statistics. A ranking of data samples in a set of observations X is normally specified by the vector of Ranks $\underline{R} = (R_1 \text{ ----} R_N)$ where Ri is the rank of the observation Xi in the sample set. For example, with a sample set X = (15, 256, 9, 4) we have R = (3, 4, 2, 1).

In a general rank test, the test statistic T is a function of the vector of ranks R.

$$T = T(R) = \sum_{i=1}^{N} C_i \cdot g(R_i)$$

Where g (·) is a function of the ranks, and Ci are coefficients that must be determined.

The data expressed in form of the scalar test statistic T is then formed into an acceptance and rejection region for the null hypothesis Ho, noise only present in the data sample as follows:

$$H_o \cdot T(R) < K$$

$$H_i \cdot T(R) \geq K$$

Where K is a constant of threshold. This study recommends the form of the nonparametric statistic and the decision procedure.

The theory of nonparametric statistics suggests that a robust detector can be obtained by formulating the speech and background noise signals in terms of a nonparametric rank test. The theoretical investigations and the limited testing suggests that this nonparametric decision approach to voiced-unvoiced-silence classification of speech should be considered for other speech processing applications where the robust issue must be addressed.

System Description

Introduction

This report describes the signal classification decision procedure, along with the theoretical justification of the nonparametric distribution model.

The system operates in the following manner:

The speech signal is low-pass filtered to 3.2 kHz (telephone bandwidth), sampled at 6.67 kHz rate and high-pass filtered at approximately 200 Hz to remove any dc or low frequency hum. The output from the high-pass filter is formatted into blocks of 100 samples (15 msec of speech data). Each block of speech is then applied to four digital filters. The output of each filter and the unfiltered speech time series are rank ordered. The rank order values are then passed to the detector or classifier algorithm. Figure 1 shows a block diagram of the detection algorithm.

Filter Selection Criteria

The digital filter bank satisfies two functional requirements for the system. First, it provides the reference noise samples for the detector. The rationale for this function will be explained in the detector description. Secondly, it filters the speech waveform to estimage the instantaneous or effective bandwidth of the speech signal.

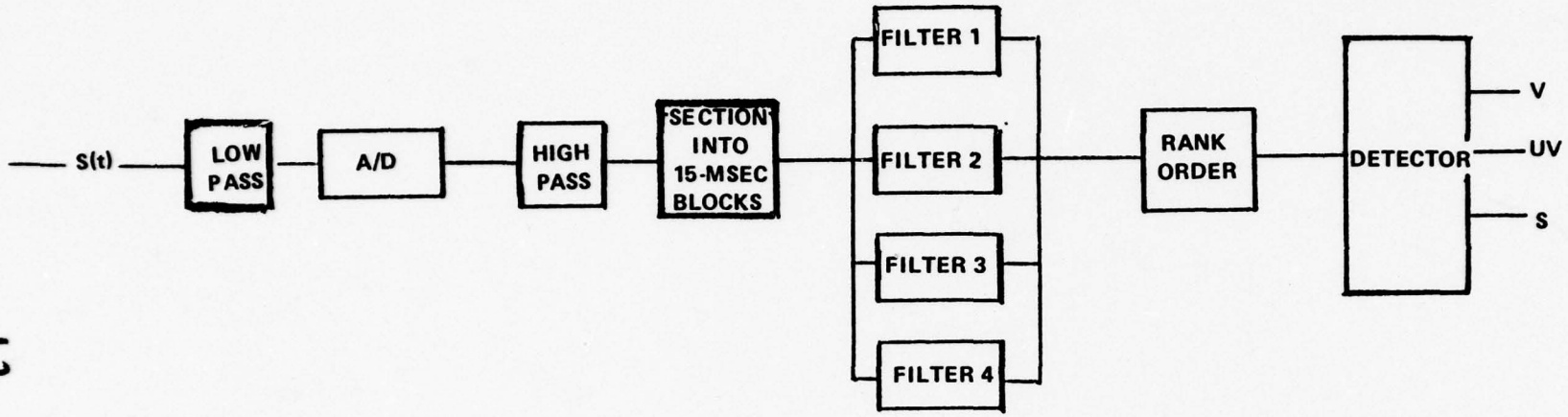


FIGURE 1. BLOCK DIAGRAM OF SIGNAL CLASSIFICATION METHOD

The filter design method is based on the work of Schafer (34) and Crochiere (35).

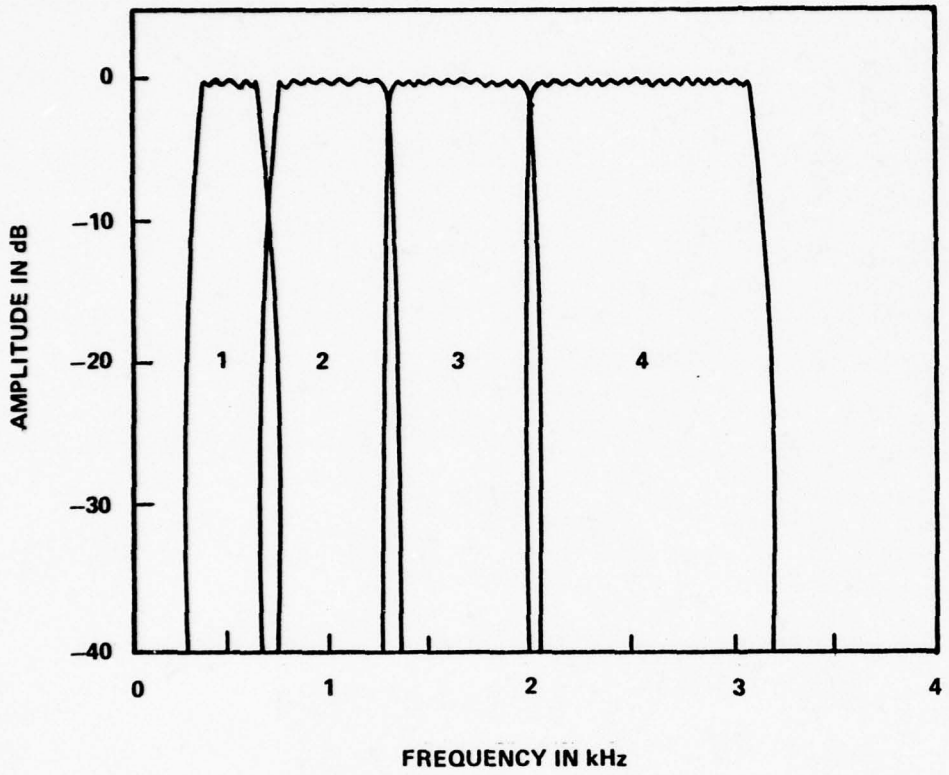
The important property achieved by this filter bank is that the sum of the individual frequency responses of the bandpass filters (composite response) lie flat with linear phase. The band-partioning is such that each sub-band contributed equally to the Articulation Index. The Articulation Index indicates, on the average, the contribution of each part of the spectrum to the overall perception of the spoken sound.

By partitioning the 200 to 3200 Hz frequency range into four equal-contribution bands, each sub-band contributing 20 percent to the AI.

This partitioning corresponds to a word intelligebility of approximately 93 *percent* (33, 34).

Figure 2 shows the partitioning of the speech spectrum into four contiguous bands.

These filters were designed using McClellan, Parks and Rabiner's program.



SUB-BAND NUMBER	FREQUENCY RANGE (Hz)
1	200 – 700
2	700 – 1310
3	1310 – 2020
4	2020 – 3200

Figure 2 PARTITIONING OF THE SPEECH SPECTRUM INTO FOUR CONTIGUOUS BANDS THAT CONTRIBUTE EQUALLY TO ARTICULATION INDEX. THE FREQUENCY RANGE IS 200 TO 3200 Hz.

Nonparametric Detector Design

Introduction

This section will present an analysis of the nonparametric decision procedures considered for experimental verification.

The purpose of the nonparametric decision procedure design is two-fold:

1. Develop and test a nonparametric detector that will reliably estimate the bandwidth occupied by the speech signal, and

2. Develop and test a nonparametric detector that will classify a given set of speech data as voiced speech, unvoiced speech or silence.

A distribution model is presented in order that the form of the distributions involved can be used to obtain a suitable decision procedure and test statistic. Using this model as a starting point, methods for estimating the noise are examined.

Data Model

To capitalize on the advantages offered by nonparametric hypothesis testing, it is sufficient to investigate the form of the expected sample distribution.

A distribution model for the speech amplitude values is required in order to be able to select a suitable test statistic.

Commeski, Palz, and Glisson (34, 36, 37, 38) and others have proposed a special form of the gamma density as amplitude probability distribution of speech

$$P_g(x) = \frac{\sqrt{K}}{2\sqrt{\pi}} \quad \frac{e^{-K|x|}}{\sqrt{|x|}}$$

where

$$\text{RMS VALUE} = \sigma_x = \frac{\sqrt{0.75}}{K}$$

A simpler approximation is the double exponential or Laplacian density

$$P_e(x) = \frac{\alpha}{2} e^{-\alpha|x|}$$

where

$$\text{RMS VALUE} = \sigma_x = \frac{\sqrt{2}}{\alpha}$$

In Figure 3, the gamma and Laplacian densities are compared with the experimentally determined density for real speech (from Palz and Glisson) (36).

The amplitude distribution is interpreted as the sum of two distribution: one distribution with a very high peak at zero amplitude

corresponds to unvoiced sounds (e.g., fricatives) and system noise, and another,that of large amplitude values corresponding to voiced sounds (e.g., vowels/semivowels, etc.).

The small value of amplitude (unvoiced and noise) can be approximated as a normal distribution, and an exponential distribution can approximate the voiced sounds.

An alternative model can be derived by formulating the detector assuming that the signal is applied to a square-law detector prior to being processed. The model to be developed is used to determine the signal power in each of the filters outputs (bandwidth BrHz). Thus, the presence of a speech signal is indicated by an increase in the average energy of the processed waveform. The signal is passed through a narrow-band bandpass filter centered at fkHz and having a bandwidth of BrHz. This filter output is then applied to a detector which consists of an absolute value process to approximate a squarer and an averager. The detector output yields an estimate of the signal power in this bandwidth.

When random Gaussian noise is applied to the input to the bandpass filter, it can be shown that the statistical properties of the output of the detector have a chi-square distribution of 2TW degrees of freedom (2, 41).

If speech is modeled as a sinusoidal signal, the density of detector output is a noncentral gamma density than can be approximated by a chi-square distribution and if modeled as flat narrow-band Gaussuan (unvoiced),

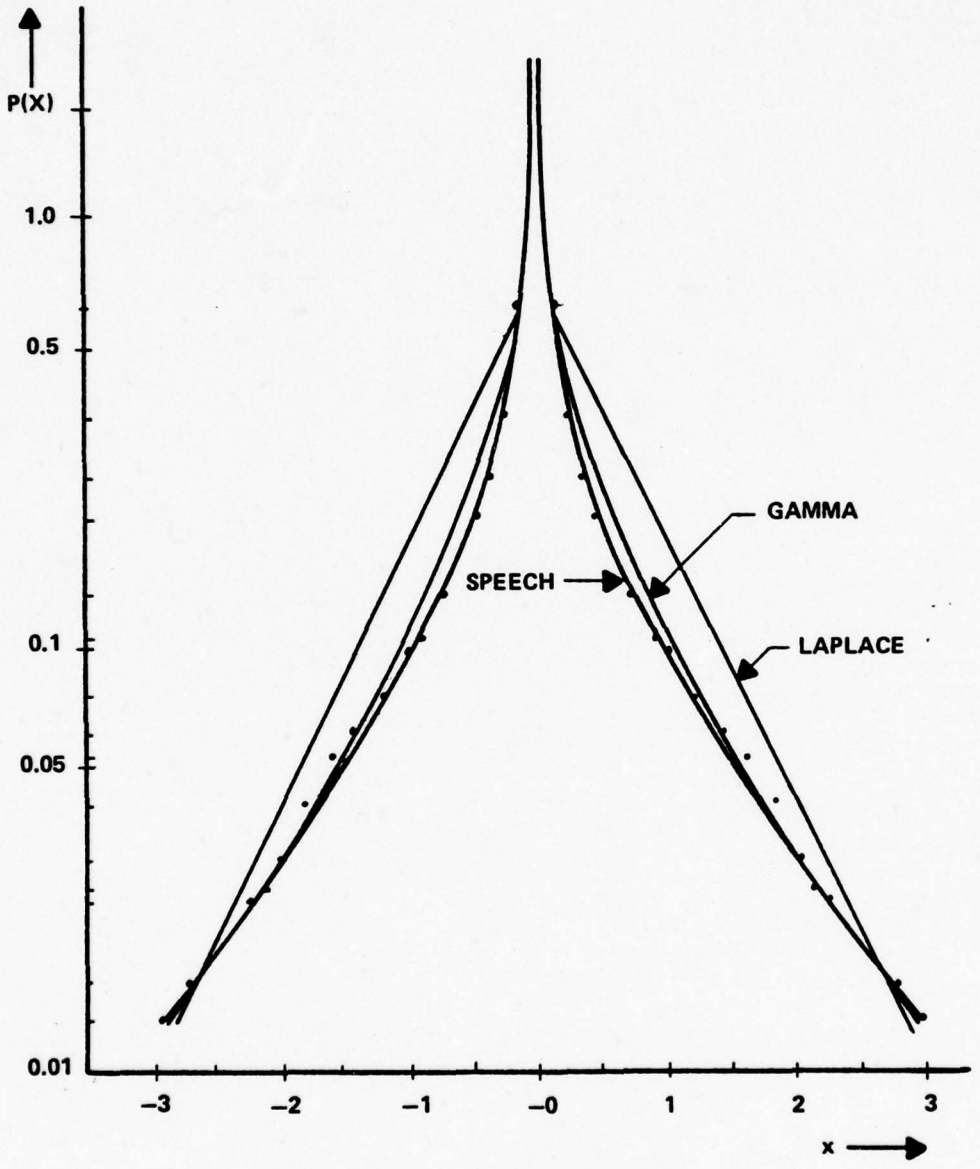


Figure 3 **REAL SPEECH AND THEORETICAL GAMMA AND LAPLACE PROBABILITY DENSITIES.**

the density is exponential (56). The aforementioned distributions are part of the gamma distribution family, and differ only in the shape parameter. The general density function for the gamma distribution is given as

$$f(x) = \frac{1}{\theta^r \Gamma(r)} x^{r-1} e^{-x/\theta}$$

$$\text{for } x \geq 0$$

The chi square distribution with ν degrees of freedom is exactly the gamma distribution with $\theta = 2$ and $r = \nu/2$.

Another special case of the gamma distribution results when r = 1; this is the exponential distribution.

Noise Estimation Procedures

The basic problem of detecting the bandwidth occupied by the speech signal can be formulated as testing the simple hypothesis H versus the composite alternative K where H and K are given by

$$H : x = V$$

$$K : x = S_i + V$$

and x is the input sample, V is zero mean Gaussian noise with unknown variance δ^2 and S_i is the speech signal to be detected.

Based on the input samples x denoted by $x = \{ x_i;\ i = 1, ---, N \}$ it is desired to develop a decision rule for determining which of the hypothesis "best" characterizes the sampled data.

Classical parametric detectors assume that f(x/H) and F(x/K) are known or than estimates the density parameters can be obtained from a training set of data.

For this investigation the assumption will be made that the noise is non-stationary. This is equivalent to assuming the noise level is unknown ahead of time by the detector and the classical approach does not apply.

Provisions for obtaining a set of reference noise samples must be incorporated into the detector design. The following method was investigated for obtaining a reference test of noise samples.

The noise samples will be derived from the output of the filter band. It is assumed that the noise process is wideband, therefore, the spectrum is approximately flat across the entire band 200 - 3200 Hz. The entire frequency spectrum of interest is partitioned into four contiguous bands (see Figure 2). Each subband is assumed to be independent.

It is assumed that sampled data x_i^j where ($J = 1, ---, 4;\ i = 1, ---, 100$) are independent for all i and j and all elements of $x^j = (x_1^j, x_2^j, ---, x_n^j)$ are identically distributed with Cdf F(s) if there is only noise in filter j, or with Foj (x) if there is speech in filter j. Woinsky (25) developed a reasonable proof that the correlation coefficient

between filters is a function of the time bandwidth product $B = 2\Delta T_f$ and for band separated by adjacent slot (j + 2) all correlations are small.

Under the null hypothesis of only noise in the filter outputs, all χ_i^j are distributed as F(x). The presence of speech in any of the x_i^j induced a Laplacian or Gaussian density which at small signal-to-noise ratio can be characterized as a scale alternative.

This result leads quite naturally to the use of two-sample statistics. The input sample data in the filter being tested form the first sample, and the pooled data from the remaining filters form the reference sample or second sample. The decision procedure proposed is a multiple decision procedure - - simultaneously test H (noise only) and upon rejection *indicates what filter output* contains speech data.

The decision procedure proposed, forms a test statistic for each filter output and compares the resulting value of the statistic with a threshold to determine what filter contains speech.

Decision Procedure (Simultaneous Test)

Under null hypothesis, H , χ_i^j is distributed as F(x), $i = 1, 2, ---, n$, $j = 1, 2, 3, 4$. Each filter output is tested for speech energy.

The two sample processing procedure is:

Let $$\chi^j = (\chi_1^j, \chi_2^j, \cdots \chi_n^j)$$

$$Y^{-j} = (x_1^1, x_2^1, \cdots x_n^1, \cdots, x_1^2, x_2^2, \cdots x_n^2, \cdots, x_1^4, \cdots\cdots x_n^4)$$

where x^j are the data to be tested and $Y-j$ are the remaining pooled data.

Let S = $(x^j, Y-j)$ where D $(\cdot,\cdot)$ is a two-sample statistic. The data are reduced to four statistics (s_1, s_2, s_3, s_4). The s_i are obviously dependent. The decision procedure is to declare there is speech in filter j if $S > \lambda\alpha$.

Where $\lambda\alpha$ is the threshold determined by the false-alarm probability as specified.

The Kruskol-Wallis One-Way Anova Test

The experiemental situation is one where K random samples have been obtained, one from each of K possible different populations, we want to test the null hypothesis that all of the populations are identical.

Sample 1	Sample 2	Sample K
X_{11}	X_{21}	$X_{K,1}$
X_{12}	X_{22}	.
.	.	.
.	.	.
.	.	.
X_{1,n_1}	X_{2,n_2}	X_{K,n_K}

The large sample approximation for the distribution is T is based onthe fact that Ri is the sum of ni random variables, and for large ni,

the Central limit theorem may be used.

$$\therefore \quad \frac{Ri - E(Ri)}{Var(Ri)} \approx N(0, 1) \qquad \text{when Ho is true}$$

and $$\left[\frac{Ri - E(Ri)}{Var(Ri)}\right]^2 \approx \text{Chi-square with one degree of freedom}$$

If the Ri were independent of each other, the distribution of the sum

$$T^1 = \sum_{i=1}^{K} \left[\frac{Ri - E(Ri)}{Var(Ri)}\right]^2 \qquad \text{Chi-square with k degrees of freedom}$$

However the sum of the Ri is K ni so there is a dependence among the Ri. If T^1 is multiplied by $\frac{N-ni}{N}$ for i = 1, 2, 3 -- K, then the result is asympotically distributed as a Chi-square with K-1 degrees of freedom.

Under Ho the savage statistic satisfies

$$E(T) = m \qquad Var(T) = \frac{m\,n}{N-1}\left(1 - \frac{1}{N}\sum_{i=1}^{N}\frac{1}{j}\right)$$

$$T^1 = \frac{(N-m)}{N} \cdot \frac{\sum_{i=1}^{K}\left[Ri - E(T)\right]^2}{\frac{m\,n}{N-1}\left(1 - \frac{1}{N}\sum_{i=1}^{N}\frac{1}{j}\right)}$$

The decision rules is declare

$$H_o : \text{if } T^1 \leq K$$
$$H_i : \text{if } T^1 > K$$

The filter with the largest rank is declared as the best estimate of the bandwidth of the speech signal. A detailed description of this nonparametric test is contained in References (42, 44, 45, 49).

Choice of a Two-Sample Statistic

The form of the distributions developed in the data model section of this proposal identifies the following two-sample statistics that will be incorporated in the speech detector.

Savage Statistic

The savage statistic was studied for this application because it is the optimum rank statistic for an exponential distribution and a scale alternative (43, 52).

Assume we want to test whether two samples differ in scale (dispersion). The procedure for the two sample problem is to combine both samples into a single ordered sample and then assign ranks to the sample values from the smallest to the largest value, without regard to which population each came from. The test statistic is the sume of ranks assigned to the values from one of the population. If the sum (test statistic) is too small, or too large, there is some indication that the values from that population tend to be smaller, or larger than the values of the other population. The null hypothesis of no difference between

populations may be rejected if the ranks associated with one sample tend to be larger than those of the other sample.

The savage statistic has the form

$$S = \sum_{i=1}^{N} AiZi$$

where

$$Zj = \begin{cases} 1 \text{ if } Xj \text{ belongs to } X1---Xm \\ 0 \text{ if } Xi \text{ belongs to } Xm+1,---Xm+n \end{cases}$$

and

$$Ai = \sum_{j=n-i+1}^{N} \frac{1}{j}$$

under Ho, the savage statistic satisfies

$$E(S) = m$$

$$VAR(s) = \frac{mn}{n-1}\left(1-\frac{1}{N}\sum_{j-1}^{N}\frac{1}{j}\right)$$

See Hajek (40) page 84 for proof.

The normal approximation since N is large will be used.

$$l \doteq \left(\frac{N-S}{\sqrt{VAR(S)}}\right)$$

accept Ho if $T \geq W\alpha$

reject Hl if $T < W\alpha$

Simplified Procedures

The rank-sum is quite complex; the procedure requires that all data in the filter be stored and also requires time consuming processing because all rank values have to be re-adjusted with each new filter output.

This can be greatly simplified using mixed statistical test.

Feustal (50) shows that Mann-Whitney statistic presented has high efficiency but requires 0 (N^2) operations to perform the ranking operation. Feustal proposed a mixed statistical test that required 0 (mn) operations. The mixed statistic operates as follows:

The N observation from each K samples are divided into p groups of m observations. An intermediate statistic on each of the observations for x samples reduces KN observations to p values. The p values are summed to form a test statistic to compare with the threshold.

The paper shows that for $m \geq 15$ negligible loss in efficiency is experienced. Woinsky extended these results to the two-sample case and through simulation confirms Feustal theoretical results.

This variation to the rank sum test will be incorporated into the experimental verification test to simplify the computational requirements.

Test Procedure

The nonparametric detector was tested on nine types of signals:

1) noise output from the output of an analog noise generator

2) background noise from the Rhyme file

3) the following words from the Rhyme file: Gob, Sue, Taunt, Nil, Boast, Jab, and Cheat

The dyagnostic Rhyme tape was supplied by Dyna Stat Inc. (72). The additive white noise tape was generated by digitizing the analog output of an analog noise generator. Both the word file and the noise file are prefiltered with a low pass filter having a 3.2 kHz cutoff frequency and is sampled at 6667. Hz.

The program ARPGEN·SAV (70) is used to calculate the SNR and the Constant C is computed by SPLUSN·SAV (70) so that specified signal-to-noise ratio test words can be created.

Using the above software programs and data files, various words with additive white noise of progressively smaller signal-to-noise ratios: 40, 30, 20, 10, and 0 dB were created and proceeded by the detector algorithms.

Preliminary Results

Evaluation Tests

Six preliminary speech tests were conducted to evaluate the speech detectors performance for five different S/N ratios: 0, 10, 20, 30, and 40 dB of wideband Gaussian noise. For each clean test word from the Rhyme file, a manual analysis was performed on each 15 msec interval to classify it as voiced, unvoiced, or silence based on visual inspection of the acoustic waveform and a phonetic interpretation of the utterance. Two independent manual classifications were made on each test word.

Gaussian noise was digitally (see test procedure) added to the clean test to produce a controlled data base having specified signal-to-noise ratios.

Error rates were computed by comparing the manual classification with the detectors classification output.

Experiment No. 1

The first test measured the correctness of assuming the zero mean Gaussian and Laplacian amplitude distribution models for noise and speech, the assumption that speech manifests itself as a scale alternative, and that the E(Ri) = 20.0.

The Mann-Whitney test (48) was performed on the noise file and

all three original word files (40 dB). The test results were not significant; the null hypothesis that mean value is essentially zero could not be rejected at the 95% confidence level. When the savage T test was used, the assumption of a scale alternative was significant and the results will be reported in the remainder of the presented results.

The mean value of the rank order statistic for the savage T measured M(Ri) = 19.97, Sx = 5.97, Sx = .56 for 300 blocks of 15 msec data. This compares to a normal approximation expected value of

$$M(Ri) = 20.00$$

$$VAR = 3.77$$

The above results did not experimentally refute the amplitude distribution model assumption of that the normal approximation for the test statistic was not valid.

Experiment No. 2

Two preliminary speech tests were conducted to evaluate the NP detector. The first test measured the accuracy of the classification algorithm when the savage T test was used to compare the amplitude distribution of speech and noise. The test method rank ordered 100 samples from each of the four filter outputs, and formed the pooled sample to estimate the noise. Two hypothesis testing procedures were employed. One approach is to test Ho by the Kruskal-Wallis test statistic and upon reject use a ranking and selection procedure to

locate the frequency location of the speech sample. The other approach employs a procedure that simultaneously tests H and upon rejection indicates what frequency component is present. This procedure is known in statistical literature under the heading of multiple-decision procedures, multiple comparison procedures, or simultaneous statistical inference.

The second test used a mixed statistical test where the absolute value of 5 samples were averaged, and 20 blocks of these averaged values were then pooled and tested as above. This method cut the ranking requirement from 100 to 20. The efficiency of this method was then compared to the classification accuracy of the full ranking algorithm.

The full ranking algorithm tests the amplitude distribution, the mixed statistical test compares the energy distributions.

For the 100 sample test, the threshold value of the classifier was set to the Chi-square approximation of 9.48.

The simultaneous test results threshold from a significance level of .05 corrected for paired comparison by $a' = \frac{a}{K\ (K-1)}$ where $a' = .0083$ and $Z = 2.39$.

For the 20 sample test, Chi-square threshold was set to 18.1. The simultaneous test threshold was set to 3.30.

Summary of Test Results

Tests of noise classification were performed on the analog noise file and the background noise in the Rhyme file.

The % correct recognition for the K-W test (20 samples) was 96%; the K-W test (100 samples) was 96.7% on the analog noise file.

The % correct recognition of the simultaneous test (20 samples) and (100 samples) was 86% and 92%.

The % correct recognition of the background noise of the Rhyme file was:

92% - K-W test (20 samples)
68% - Simultaneous test (20 samples)
93% - K-W test (100 samples)
85% - Simultaneous test (100 samples)

The % recognition test the seven words from the Rhyme file was performed using the simplified procedure (sum 5 samples, rank the 20 sums).

Table 1 summarizes the overall recognition rate as a function of S/N ratio of the simultaneous decision procedure for all the test utterances.

SNR	39	30	20	10	0
Silence	58	88	96	94	93
Voiced	94	95	88	74	53
Unvoiced	95	96	84	59	35
Total %	82	93	89	75	60

TABLE 1

Recognition Rate for the Simultaneous Decision Procedure for all Seven Words

(20 samples)

Silence 85.8% recognition overall
Voiced 80.8% recognition overall
Unvoiced 73.8% recognition overall

Table 2 summarizes the recognition rate for each word as a function of S/N ratios for the K-W decision procedure.

% Recognition	Silence					Voiced					Unvoiced				
Word S/N	39	30	20	10	0	39	30	20	10	0	39	30	20	10	0
Gob	-	-	-	-	-	95	93	77	46	18	-	-	-	-	-
Sue	88		100	100	100	94		100	100	91	71	-	17	1	1
Taunt	90	100	100	100	100	95	95	95	78	45	100	100	0	0	0
Nil	38	100	100	100	100	100	100	89	46	41	-	-	-	-	-
Boast	78	100	100	100	100	100	100	94	72	56	100	100	33	0	0
Jab	100	100	100	100	100	84	78	57	38	14	50	100	75	50	25
Cheat	91	100	100	100	100	88	88	88	82	82	86	77	71	71	29
Average %	80	100	100	100	100	93.7	92	85	66	49	81	94	49	40	18

TABLE 2 - Recognition Rate for the K-W Decision Procedure
(20 Samples)

Silence Overall Recognition Rate = 96%
Voiced Overall Recognition Rate = 77%
Unvoiced Overall Recognition Rate = 56%

Conclusion

A nonparametric statistical detector for recognizing speech has been described and implemented. Preliminary results of limited testing show that the detector performs as well as the pattern recognition approach reported in the literature (18, 19, 20, 21, 22, 23). In limited testing, the classifier performed with a misclassification rate of less than 5% with thresholds calculated from theory. The desirable feature of this detection or classification scheme are that it does not require a training set of data or apriori information of the statistical parameter of speech or noise.

The manual analysis of the speech waveform contained in segments in which uncertain intervals occurred. These uncertain intervals were mostly at stop gap in words such as "boast" and "taunt". The interval corresponding to the stop gap before the unvoiced T, amplitude and frequency characteristics resembled noise but did not contain the wideband noise characteristic.

If the segments were suppressed and the word was acoustically transcripted, no loss of intelligibility occurred. For this reason, the segments were classified as noise.

This classification technique resulted in an increase in recognition rate between the 39 dB and 30 dB signal-to-noise test. In effect, the addition of noise pre-whited the low frequency component of the recording noise and aided the classification algorithm. This uncertainty also occurred in the word "cheat". In this word during

the voicing interval, eight transition blocks occurred. These transitions were classified as errors in the algorithm detection of voicing.

The classification of the unvoiced "T" inthe word "boast" and "taunt" was correct until the added noise obscured the "T" sound. This was 10 dB for the word "taunt" and 20 dB for the word "boast". The initial "S" in "sue" is covered by noise at 20 dB. The classification algorithm decision recognition rate is calculated by using the original 39 dB clean speech utterance as the reference. If a manual reclassification was accomplished at each level of signal-to-noise ratio, the resulting algorithm classification compared with one or two percent of the same recognition rate as the original clean speech classification. The 20 sample test compared favorably with the 100 sample test when the threshold was adjusted to compensate for the change in number of degrees of freedom; and the loss of efficiency because of using the mixed statistical decision procedure. The 20 sample test will be used for the remaining analysis in this research. The simultaneous test procedure will be used instead of the K-W decision procedure because it is more sensitive to the detection of the unvoiced interval of speech. The loss of efficiency for recognizing the silence interval will be investigated and an alternate decision algorithm implemented to correct this deficiency.

Future Research

The results of the tests indicate that sufficient increases in false alarm rates are experienced with the mixed statistic simplified algorithm. The reasons for the negative results are at present not completely understood. Loss of efficiency is predicted from theory, but the correction to the threshold setting to offset this disadvantage isnot readily apparent. A more detailed comparison of the full rank algorithm (100 samples) and the mixed statistic algorithm will be studied. Alternate decision algorithms proposed in the theoretical description will also be tested.

The conditional rank test approach proposed by Kassam (68), was tried on a limited set of data and looks as though this technique can improve the silence false alarm rate. More extensive testing of this approach will be undertaken. A modified level test where the two sample test updates a fixed threshold that regulates the false alarm probability and a second threshold that is adaptive and estimates the standard deviation of the pooled samples noise estimate to compensate for loss of efficiency of unvoiced decision at low S/N will be tested. Preliminary test of this concept resulted in 100% recognition of the unvoiced segment of the word "Taunt" down to S/N of 0 dB.

A study of the effects of incorporating these techniques into the proposed detection algorithm will be reported in the next semi-annual report.

REFERENCES

(1) H. Mudema and M. G. Schachtman, "TASI Quality-Effect of Speech Detection and Interpolation," BSTJ, July 1962.

(2) Harry Urkowitz, "Energy Detection of Unknown Determinatic Signals", Proceedings of IEEE, Vol. 55, No. 4, April 1967.

(3) Thomas E. Eger and John E. Whelchel, Jr., "Variable Rate Digital Voice Communications", EASCON '74

(4) Whelchel, J. E., R. H. Strand, Jr., T. E. Eger, G. F. Mayifskie, "Design of an Adaptive PCM Speech Transmition System" 1974 ICC Conference Proceedings

(5) S. J. Campanella and H. S. Seyerhoud, "Digital Speech Interpolation for Telephone Communications," EASCON '74

(6) Thomas E. Eger and S. J. Campanella, "Study of an Adaptive Bandwidth PCM Voice Communication System" 1974 ICC Conference Proceedings

(7) J. A. Sciulli and S. J. Campanella, "A Speed Predictive Encoding Communication System for Multichannel Telephony" IEEE Trans on Communication, Vol. COM-21, NO7, July 1973

(8) L. H. Rosenthal, L. R. Rabiner, R. W. Schafer, P. Cummiskey, J. L. Flanagen, "A Multiline Computer Voice Response System Utilizing ADPCM Coded Speech", IEEE Trans on Acoutics, Speech, and Signal Processing, Vol. ASSP-22, No.5, October 1974.

(9) E. Lyghouris, I. Poretti, and G. Monti, "Speech Interpolation in Digital Transmission Systems". IEEE Trans on Communication, Vol. COM-22, No. 9, September 1974.

(10) H. F. Silvermon and N. R. Dixon, "A Parametrically Controlled Spectral Analysis System for Speech", IEEE TRANS on Acoustics, Speech, and Signal Processing, Vol. ASSP-22, No. 5, October 1974

(11) H. F. Silvermon and N. R. Dixon, "Transfer Characteristics Estimation for Speech Via Multirate Evaluation", EASCON '75 Conference Proceedings

(12) R. W. Schafer and L. R. Rabiner, "Digital Representations of Speech, Signals", Proceedings of the IEEE, Vol. 63, No. 5, April 1975

(13) S. Seniff, "A Real-Time Digital Telephone Simulation on the Lincoln Digital Voice Terminal", MIT Lincoln Digital Voice Terminal, Technical Note 1975-6, 30 December 1975, ESD-TR-75-326

(14) B. Gold, "Robust Speech Processing", MIT, Lincoln Laboratory, Technical Note 1976-6, January 1976

(15) L. R. Rabiner, et all, "Special Issue on Man-Machine Communications by Voice", Proceedings of the IEEE, Vol. 64, No. 4, April 1976

(16) S. G. Pitnoda and B. J. Rikienie, "A Digital Conference Circuit for an Instant Speaker Algorithm", IEEE TRANS on Communication Technology, Vol. COM-19, No. 6, December 1971.

(17) E. Farello, "A Novel Digital Speech Detector from Improving Effective Satellite Capacity", IEEE TRANS on Communication, February 1972.

(18) L. R. Rabiner and M. R. Sambur, "Some Preliminary Experiments in the Recognition of Connected Digits", IEEE TRANS on Acoustics, Speech, and Signal Processing, Vol. ASSP-24, April 1976.

(19) L. R. Rabiner, and M. R. Sambur, "Speaker Independent Recognition of Connected Digits", IEEE International Conf. on Acoustics, Speech, and Signal Processing, Conference Record, April 1976.

(20) R. W. Schafer, K. Jackson, J. J. Dubnowski, and L. R. Rabiner, "Detecting the Presence of Speech Using SDPCM Coding", IEEE TRANS on Communication, May 1976.

(21) Bishnu S. Atal and L. R. Rabiner, "A Pattern Recognition Approach to Voice-Unvoiced-Silence Classification with Application to Speech Recognition", IEE TRANS on Acoustics, Speech, and Signal Processing, Vol. ASSP-24, No. 3, June 1976.

(22) L. R. Rabiner, C. E. Schmidt, and B. S. Atal, "Evaluation of a Statistical Approach to Voiced-Unvoiced-Silence Analysis for Telephone Quality Speech", BSTJ, March 1977.

(23) L. R. Rabiner and M. R. Sambur, "Voiced-Unvoiced Detection Using the ITAKURA LPC Distance Measure," 1977 IEEE International Conference on Acoustics, Speech and Signal Processing, Hartford, Conn., May 1977

(24) J. L. Dubnowski, R. W. Schafer, and L. R. Rabiner, "Real-time Digital Hardware Pitch Detector", IEEE TRANS on Acoustics, Speech and Signal Processing, Vol. ASSP-24, No. 1, February 1976.

(25) J. A. Jankowski, Jr., "A New Digital Voice-Activated Switch", Comsat Technical Review, Vol. 43, No. 4, Spring 1976.

(26) W. J. Hess, "A Pitch-Synchronous Digital Feature Extraction System for Phonemic Recognition of Speech", IEEE TRANS on Acoustics, Speech and Signal Processing, Vol. ASSP-34, No. 1, February 1976.

(27) L. J. Siegel and K. Steiglits, "A Pattern Classification Algorithm for the Voiced/Un-voiced Decision", 1977 IEEE International Conference on Acoustices, Speech and Signal Processing, Hartford, Conn., May 1977

(28) F. Daabond and J. P. Adoul, "Parametric Segmentation of Speech into Voiced-Unvoiced-Silence Intervals", _____, _____, _____.

(29) Jean-Preme Adoul and D. Prodelles, "On line Speech/DATA-modem identification for Telephone Network", _____, _____, _____.

(30) R. J. McAulay, "Optumum Classification of Voice Speech, Unvoiced Speech and Silence in the Presence of Noise and Interference", MIT Lincoln Laboratory, Technical Note 1967-77, 3 June 1976.

(31) R. J. McAulay, "Optimum Speech Classification and Its Application to Adaptive Noise Cancellation", Lincoln Laboratory, LEX, MASS, Technical Note 1976-39, 9 November 1976

(32) P. G. Drago, A. M. Molinari and F. C. Vagliani, "Digital Dynamic Speech Detectors", IEEE TRANS on Communications, Vol-Com-26, No. 1, January 1978.

(33) R. E. Crochiere, S. A. Webber, and J. L. Flanagan, "Digital Coding of Speech in Sub-bands", BSTJ, Vol. 55, No. 8, October 1976

(34) R. W. Schafer, L. R. Rabiner, and O. Herrmann, "FIR Digital Filter Banks for Speech Analysis", BSTJ, Vol. 54, No. 3, March 1975

(35) R. E. Crochiere and M. R. Sambur, "A Variable Band Coding Scheme for Speech Encoding at 4.8 kb/s", IEEE, International Conference on Acoustics, Speech and Signal Processing, Hartford, Conn., May 1977

(36) M. D. Paez and T. H. Glisson, "Minimum Mean-Squared-Error Quantization in Speech PCM and DPCM Systems", IEEE TRANS on Communications, April 1972.

(37) E. H. Rothauser, "Speech in Digital Communications Systems", IEEE TRANS on Audio and Electroacoustics, Vol. AV-21, No. 1, February 1973.

(38) D. Chan and R. W. Donaldson, "Subjective Evaluation of Pre-and Post-fitting in PAM, PCM, and DPCM Voice Communication Systems", IEEE TRANS on Communication Technology, Vol. Com-19, No. 5, October 1971

(39) Barlow, Bartholomew, Bremner, and Brunk, Statistical Inference Under Order Restrictions, New York: Wiley, 1972

(40) Jaroslav Hajek, A Course in Nonparametric Statistics, San Francisco: Holden-Day, 1969

(41) Bendat, Principles and Applicaitons of Random Noise Theory, New York: Wiley, 1958.

(42) W. J. Conover, Practical Nonparametric Statistics, New York: Wiley, 1971

(43) E. L. Lelmann, Nonparametrics: Statistical Methods Based on Ranks, San Francisco: Holden-Day, 1975

(44) M. Hollander, and D. A. Wolfe, Nonparametric Statistical Methods, New York: Wiley, 1973

(45) Gibbons, Nonparametric Statistical Inference, New York: McGraw-Hill, 1971

(46) H. A. David, Order Statistics, New York: Wiley, 1970

(47) R. G. Miller, Simultaneous Statistical Inference, New York: McGraw-Hill, 1966

(48) J. D. Gibson and J. L. Melsa, Introduction to Nonparametric Detection with Applications, New York: Academic Press, 1975

(49) G. E. Noether, Introduction to Statistics, Boston: Houghton Miffline Company, Second Edition, 1976

(50) E. A. Feustal, and L. D. Davisson, "The Asymptatic Relative Efficiency of Mixed Statistical TEST", IEEE TRANS on Information Theory, Vol. IT-13, No. 3, September 1968

(51) _____, "On the Efficacy of Mixed Locally-Most-Powerful One-and Two-Sample Rank Tests", IEEE TRANS on Information Theory, Vol, IT-13, No. 3, September 1968

(52) I. R. Savage, "Contributions to the Theory of Rank Order Statistics--the Two-Sample Case", Ann Math Statist, Vol. 27, 1956

(53) S. Siegel, and J. W. Tukey, "A Nonparametric Sum of Ranks Procedure For Relative Spread in Unpaired Samples", Ann Statist Assn Jour, September 1960

(54) P. Papantoni-Kazakos, "Small-SampleEfficiencies of Rank Tests", IEEE TRANS on Information Theory, Vol, IT-21, No. 2, March 1975

(55) S. A. Kassan and J. B. Thomas, "Generalizations of the Sign Detector Based on Conditional Tests", IEEE TRANS on Communications, Vol. COM-24 No. 51, May 1976

(56) M. W. Woinsky, "Nonparametric Detection Using Spectral Data", IEEE TRANS on Information Theory, Vol. IT-18, No. 1, January 1972

(57) Yau-Chau Ching, and Ludwik Kurz, "Nonparametric Detectors Based on m-Interval Partitioning", IEEE TRANS on Information Theory, Vol. IT-18, No. 2, March 1972.

(58) R. D. Martin, "Robust Estimation of Signal Amplitude", IEEE TRANS on Information Theory, Vol. IT-18, No. 5, September 1972.

(59) R. D. Martin and S. C. Schwartz, "Robust Detection of a Known Signal in Nearby Gaussian Noise", IEEE TRANS on Information Theory, Vol. IT-17, No. 1, January 1971.

(60) J. B. Thomas, "Nonparametric Detection", IEEE Proceedings, Vol. 58, No. 5, May 1970

(61) M. N. Woinsky, "A Composite Nonparametric Test for a Scale Slippage Alternative", Ann Math. Statist., Vol. 43, No. 1.

(62) R. F. Daly and C. K. Rushforth, "Nonparametric Detection of a Signal of Known Form in Additive Noise", IEEE TRANS on Information Theory, Vol. IT-11, Jan. 1965.

(63) H. L. Groginsky, L. R. Wilson, and David Middleton, "Adaptive Detection of Statistical Signals in Noise", IEEE TRANS on Information Theory, Vol. IT-12, No. 3, July 1966.

(64) M. B. Sirvanci and S. S. Wolff, "Nonparametric Detection with Autoregressive Data", IEEE TRANS on Information Theory, Vol. IT-22, No. 6, November 1976.

(65) Yves, Lagae, "A Combination of Wilcoxon's and Ansari-Bradley's Statistics", Beometrika, Vol. 58, No. 1, 1971

(66) R. Elsner, W. K. Endners, H. Mongold, et al, "Recent Progress in Digital Processing of Speech", IEEE TRANS on Communications, Vol. COM-22, No. 9, September 1974.

(67) Olive Jean Dunn, "Multiple Comparisons Using Rank Sums", TECHNOMETRICS, Vol. 6, No. 3, August, 1964

(68) S. A. Kassam, "A Conditional Rank Test for Nonparametric Detection", IEEE TRANS on Information Theory, Vol. IT-23, No. 3, May 1977

(69) B. Gold, "Digital Speech Networks", Proceedings of the IEEE, Vol. 65, No. 12, December 1977

(70) S. F. Boll, et al, "Noise Suppression Methods for Robust Speech Processing", Semi-Annual Technical Report, UTEC-CSC-77-202, Computer Science Dept. University of Utah, April 1977

(71) Win C., Lin and C. F. Chon, "An Isolated Word Recognition System Based on Acoustic-Phonetic Analysis and Statistical Pattern Recognition", IEEE International Conference on Acoustics, Speech, and Signal Processing, Hartford, Conn., May 1977.

(72) William D. Voiers, Alan D. Sharpley, and Carl H. Hehmosoth, Research on Diagnostic Evaluation of Speech Intelligibility, Final Report AFSC Contract No. F19628-70-C-0182 1973

THE CONSTANT-Q TRANSFORM

Jim Kajiya

1. Introduction

In signal processing the use of the Fourier transform has enjoyed singular success not only as a practical tool of unmatched power and rich application but also as a theoretical viewpoint imparting simplifying insight. Unfortunately a problem arises in that the Fourier integral transform:

$$\hat{f}(\omega) = \int_{-\infty}^{\infty} f(t)e^{-i\omega t}dt$$

is impossible to calculate on a computer, thus we need to construct approximations in order to make calculation of the Fourier transform effective. First we usually approximate the integral by a sum i.e. time-sample the input signal:

$$\hat{f}(\omega) = \sum_{n=-\infty}^{\infty} f(nT)e^{-i\omega nT}$$

Even with this simplification an infinite sum remains, therefore in order to make the calculation in a reasonable amount of time we must truncate to a finite number of terms:

$$\hat{f}(\omega) = \sum_{n=-N/2}^{+N/2} f(nT)e^{-i\omega nT}$$

This second approximation is equivalent to truncating the input signal to a finite record length. We are now confronted with a problem: How do we choose our approximations wisely in order to maintain a reasonable accuracy in our calculations? For time-sampling, the familiar Shannon sampling theorem provides a guide so that we can make wise decisions concerning this approximation. The second approximation, truncating the input signal or "windowing," has been handled by engineers on an ad hoc basis. There has been no analog of the Shannon sampling theorem to help us decide how long to make our window for any given application. Indeed, in most applications it is apparent that a single optimum window length doesn't exist: for example in speech processing it seems that medium length windows are both too short in some respects and too long in others.

The balance of this chapter will develop techniques to circumvent the need for choosing a window size. In the next section, we make a quick review of the current fixed window length technique; which (for reasons that will become evident) we will call constant bandwidth technology. Second, a broader view of the signal processing discipline is taken so that we may understand some fundamental effects of the constant bandwidth technology approximation procedure. This fundamental discussion will show why the constant

bandwidth technology is for speech processing a poor approximation to the ideal Fourier transform. The third section uses the fundamental discussion to generate a good approximation to the Fourier transform which we call constant-Q technology. Fourth, follows a list of applications of the constant-Q technology.

2. Constant Bandwidth Technology

Since sampled speech is a quasi-infinite sequence of points most speech processing techniques such as LPC, Homomorphic, SABRE, etc. segment this sequence into a series of (possibly overlapping) finite length records called windows. Implicit within all these segmenting techniques is a popular operation known as the short-time spectrum. The (proto-sampled) short-time spectrum is expressed as follows:

$$\hat{f}(\omega,t) = \int_{-\infty}^{\infty} f(\tau)\, h(t-\tau) e^{-i\omega\tau}\, d\tau$$

where $f(\tau)$ is the input speech signal, $h(\tau)$ is a compact support (finite length) "windowing" function $\hat{f}(\omega,t)$ is the output function with ω indexing the frequency and t indexing the center of the window. This process has been known for a long time [1]. To achieve computational effectiveness we need only to sample the input signal and convert the integral to a sum. Since the window function has finite length this sum is automatically finite hence calculatable.

Interesting issues arise when one asks for an "inverse" to this transform [2]. We can write the inverse as follows:

$$f(t) = \frac{1}{\langle g,h \rangle} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(t-\tau)\hat{f}(\omega,\tau) e^{i\omega t}\, d\omega d\tau$$

where $\langle g,h \rangle$ is the Hilbert space L inner product, g is a suitable reconstruction function. (For a guide to choosing g see [2].

The following manipulation of the short time spectrum is important not only for theoretical insight but also for practical applications. We rewrite $\hat{f}(\omega,t)$ as follows:

$$\hat{f}(\omega,t) = e^{-i\omega t} e^{i\omega t} \int_{-\infty}^{\infty} f(\tau)h(t-\tau)e^{-i\omega\tau}\, d\tau$$

$$= e^{-i\omega t} \int_{-\infty}^{\infty} f(\tau)h(t-\tau)e^{i\omega(t-\tau)}\, d\tau$$

$$= e^{-i\omega t} \left\{ f(t) * \left[h(t)e^{i\omega t} \right] \right\}$$

The last expression shows that for fixed ω, $\hat{f}(\omega,t)$ is the baseband demodulated output of a bandpass filter with center frequency ω. Each filter has a bandwidth determined only by h and is independent of its center frequency. Hence the appellation constant bandwidth.

Note that in order to present $\hat{f}(\omega,t)$ we need a two-dimensional display (an image). This process then allows two-dimensional processing techniques to be used in speech processing.

For the above and many more reasons $\hat{f}(\omega,t)$ is coming out of its implicit role in the old algorithms to enjoy an explicit status in new algorithms. Recently the popularity of the short-time spectrum has been growing [3,4,5,6].

3. Perception and Signal Processing

In order to understand the fundamental aspects of windowing it is necessary to gain a broader perspective of the relationship between perception and signal processing. The question we address is: Why does the Fourier Transform (instead of the Mellin, Hankel, etc.) enjoy such singular success in the signal processing discipline? The answer is: Because the Fourier transform is "matched" with the manner in which the signals to be processed are produced and consumed. Specifically, a speech signal is produced by convolving a glottal wave with the vocal tract impulse response. In turn, a speech signal is consumed by the human auditory perception system. This criterion of *matching* a transform with the signal production and consumption mechanism is a natural one for engineers. The two-dimensional time-frequency spectrograms which display the $\hat{f}(\omega,t)$ of the short-time spectrum correspond roughly to processes known to occur in the Human Auditory System; this explains in part the appeal of $\hat{f}(\omega,t)$ and its growing popularity. We shall ignore the matching between the transform and the signal production mechanism and concentrate on the relationship between the transform and the signal consumption mechanism. To explain what we mean

precisely by "matching" it is necessary to talk about the theory of Lie group representations. This would take us too far afield for the present exposition so we will attempt to paraphrase results. A particularly powerful observation one can make about any given system is the symmetry operations one can perform on that system. For example, a symmetry operation consonant with the auditory system is time translation, since perception of a speech waveform is unaffected by a pure time delay. However, the auditory system is not subject to time reversal symmetry: reversal of a speech waveform completely destroys intelligibility. The set of symmetry operations can be concatenated and inverted, i.e. they form a group. Using Lie Group Theory one can deduce all transforms related to that group. These "symmetric" transforms are called equivariant or intertwining operators. For example, the Fourier transform can be deduced solely on the basis of the group of time translations above. This then is what we mean by a matched transform and system: The symmetry operations must be the same; as in the case of the time translation group corresponding to both the auditory system and the Fourier transform.

One can raise the question: Are there further symmetries corresponding to the auditory system? There are several clues from psychophysical studies. As we mentioned above, the auditory system does process

speech into a time-frequency format. However, experiments don't seem to uncover any basic window. For example, under a constant bandwidth (fixed window) regime the time resolution at all frequencies remains the same. But the auditory system analyses high frequencies with much finer time resolution than low frequencies. Also modeling the Auditory system as a bank of tuned bandpass filters one must use (instead of constant bandwidth filters) constant-Q filters i.e. filters with bandwidths a fixed percentage of their center frequencies. These properties suggest that a symmetry group of the auditory system is the group of time scaling and shifting--the at+b group.

Simple calculus reveals that the Fourier transform reacts gracefully to the at+b group: the Fourier transform not only has a simple time translation property but also a simple time scaling property. The fundamental observation about the windowing approximation is that it destroys the scaling property enjoyed by the Fourier transform In fact the symmetry group corresponding to the constant bandwidth technology is time translation and sinusoidal modulation, i.e. frequency translation. That frequency translation is not a symmetry of the Auditory system is readily apparent to anyone who has tried to understand a mistuned single sideband speech signal.

4. Constant-Q Technology

This section discusses a time-frequency transform that corresponds to the at+b group. From Lie group theory one obtains the following intertwinning operator:

$$f(\tau,t) = \int_{-\infty}^{\infty} k\left(\frac{\tau-\xi}{\tau-\xi}\right)\frac{1}{(t-\xi)}\ f(\xi)\ d\xi$$

To obtain a time-frequency analog we Fourier transform along the first variable and combine expressions to obtain:

*) $$\hat{f}(\omega,t) = \int_{-\infty}^{\infty} h(\omega(t-\xi))\ f(\xi)\ e^{-i\omega\xi}\ d\xi$$

where h is the Fourier transform of k and serves as the window function. The matter of an inverse is more delicate than for the constant bandwidth case. An inverse can be shown to be:

**) $$f(t) = \lim_{\varepsilon\to 0}\frac{1}{-2k\ \log\varepsilon}\int_{|\omega|>\varepsilon}\int_{-\infty}^{\infty} f(\omega,\tau)g(\omega(t-\tau))e^{i\omega\tau}|\omega|\ d\tau d\omega$$

where $k = \int_{-\infty}^{\infty} g(u)du \cdot \int_{-\infty}^{\infty} h(u)e^{iu}du$.

A number of notable features of *) and **) should be mentioned. First, the window length of h changes for

each frequency ω so that no fixed window length need be chosen. Second, the time resolution for the higher frequencies is sharper than for the lower frequencies; thus the constant-Q technology mimics the time resolution properties of the auditory system. Third, a manipulation similar to the constant bandwidth case reveals a filter bank analogy. But in this case instead of identical bandwidth filters the bandpass filters increase in bandwidth as their center frequencies increase, thus maintaining a fixed ratio of bandwidth to center frequency (Q). Finally the constant-Q technology preserves symmetry under the at+b group (since it was designed to do so) thus satisfying our matching criterion between transform and perceptual system.

5. Applications

Almost all speech processing algorithms implicitly contain some version of the constant bandwidth technology. To apply the constant-Q technology one must make explicit $\hat{f}(\omega,t)$ and replace the constant bandwidth with constant-Q. LPC is an excellent example: When viewed as a spectral approximation technique it is easily seen that $\hat{f}(\omega,t)$ is the object that is actually being approximated by LPC. Recently Makhoul [7] has found that an ad hoc smoothing of the higher frequencies results in an elimination of the "buzziness" in LPC speech. The constant-Q technology supplies a built-in smoothing so that this very important quality issue in LPC is automatically solved by constant-Q. This should not be surprising given the fundamental matching criterion between transform and auditory system. Further applications arise from an AGC/Blind deconvolution algorithm. Also the noise reduction methods mentioned elsewhere in this report can also be converted from constant bandwidth to constant-Q technology. Since the present algorithms utilize windowing it is not surprising, given the fundamental criterion, that disturbing quality degradations are inflicted upon material to be restored.

Finally it should be noted at it is not known whether clever digital implementations of the constant-Q algorithm exist. Current digital versions run quite slowly since they are implemented via fast convolution as a filter bank. However, the fact that the bandwidth of the filters increase with center frequency allows approximately a tenfold decrease in the number of filters required by the old constant bandwidth technology. In addition, although clever digital implementations are unknown; straightforward CCD implementations class among the easiest and most natural applications of CCD's. A single CCD bandpass prototype can be used for all channels simply by fixing the clock rate to be some multiple of the center frequency.

References

1. J. L. Flanagan and R. M. Golden "Phase Vocoder," BSTJ V. 45,, pp.1493-1509, Nov. 1966.

2. J. Kajiya and J. Youngberg "On the Inverse Short Time Spectrum," to appear.

3. M. R. Portnoff, "Implementation of the Digital Phase Vocoder using the Fast Fourier Transform," IEEE Trans. Acoust.., Speech, Signal Processing, Vol. ASSP. 24, pp. 243-246, June 1976.

4. M. Callahan "Acoustic Signal Processing based on the Short-Time Spectrum", U of U Computer Science Technical Report, UTEC-CSc-76-209, March 1976.

5. J. B. Allen, "Short Term Spectral Analysis, Synthesis, and Modification by Discrete Fourier Transform," IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-25, pp 235-238, June 1977.

1. J. B Allen, and L. Rabiner, "A Unified Approach to Short-Time Fourier Analysis and Synthesis," Proc, of IEEE, Vol. 65, pp. 1558-1564, Nov. 1977.

7. J. Makhoul, et. al., "A Mixed Source Model for Speech Compression Synthesis" ICASSP 78 Conference Proceedings, Tulsa, Ok., April 1978.